REPORT RESUMES

ED 003 630 56

INCREASING LONG-TERM RETENTION OF KNOWLEDGE, EXPERIMENT 1 AND 2--METHODS OF INSTRUCTION FOR STUDENTS OF DIFFERENT ABILITY LEVELS.

BY- HAMILTON, NANCY RUSSELL
AMERICAN INST. FOR RESEARCH IN BEHAVIORAL SCIENCES
REPORT NUMBER NDEA-VIIA-1049 PUB DATE JUL 65
REPORT NUMBER AIR-D49-7-65-FR
GRANT OEG-7-48-7670-204
EDRS PRICE MF-$0.27 HC-$7.36 184P.

DESCRIPTORS- *AUTOINSTRUCTIONAL METHODS, AUTOINSTRUCTIONAL PROGRAMS, *PROGRAMED INSTRUCTION, PACING, *GRADE 8, *LEARNING, LEARNING PROCESSES, *OVERT RESPONSE, MEASUREMENT TECHNIQUES, RETENTION STUDIES, ABILITY, PALO ALTO, CALIFORNIA

AN AUTOINSTRUCTIONAL PROGRAM WITH READING-ONLY RESPONSE VERSUS OVERT RESPONSE WAS ADMINISTERED 1 HOUR EACH DAY FOR 3 WEEKS TO SIX CLASSES OF EIGHTH-GRADE STUDENTS. REVIEW PROGRAMS WERE ALSO GIVEN UNDER THE TWO RESPONSE TYPES AND UNDER CONDITIONS OF EITHER SPACED OR MASSED PRACTICE. SEVERAL MEASURING INSTRUMENTS WERE USED TO ASSESS LEARNING FROM SEPARATE ASPECTS OF THE SUBJECT MATTER--GENERAL, CHRONOLOGICAL, OR CONCEPTUAL. MEASUREMENTS WERE MADE FOR PREACQUISITION, POST-ACQUISITION, POST-REVIEW, AND DELAYED-RETENTION AFTER A SUMMER VACATION. ALL ANALYSES WERE MADE IN TERMS OF THREE LEVELS OF STUDENT ABILITY. FOR GAIN AND RETENTION OF CHRONOLOGICAL INFORMATION, THE OVERT RESPONSE WAS OF GREATER VALUE THAN READING-ONLY. ON THE GENERAL MATERIALS, HIGH ABILITY STUDENTS PROFITED FROM READING-ONLY AS MUCH AS FROM OVERT RESPONSE. FOR MIDDLE ABILITY STUDENTS AND GENERAL MATERIALS, A SIGNIFICANTLY FAVORABLE EFFECT WAS FOUND FROM HAVING A DIFFERENT MODE OF RESPONSE FOR REVIEW THAN WAS EMPLOYED IN INITIAL INSTRUCTION. LOW ABILITY STUDENTS APPEARED TO HAVE RETAINED LITTLE OR NOTHING. NO CLEAR VALUE WAS DEMONSTRATED EITHER FOR OR AGAINST SCHEDULES OF PRACTICE. (JC)

ED003630

AIR-D49-7/65-FR

1049

INCREASING LONG-TERM RETENTION OF KNOWLEDGE:

EFFECTS ON LEARNING AND RETENTION FOR STUDENTS OF DIFFERENT ABILITY LEVELS

EXPERIMENT 1: READING RESPONSE VERSUS OVERT RESPONSE PROGRAMS FOR INITIAL INSTRUCTION AND REVIEW

EXPERIMENT 2: INSTRUCTION FROM SMALL STEP VERSUS LARGE STEP PROGRAMS WITH IMMEDIATE VERSUS DELAYED REVIEW PROCEDURES

Nancy Russell Hamilton

July 1965

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Office of Education
Grant Number Title VII: 7-48-7670-204

Final Report

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

AMERICAN INSTITUTE FOR RESEARCH/PALO ALTO

AMERICAN INSTITUTE FOR RESEARCH
mailing address: Station A, P.O. Box 11487, Palo Alto, Calif.
Telephone 321-2130 (Area Code 415)
office location: 1791 Arastradero Road, Palo Alto, California

AIR-D49-6/65-FR-I

INCREASING LONG TERM RETENTION OF KNOWLEDGE,

Experiment ~~I~~ 1--Reading Response versus Overt Response Programs for Initial Instruction and Review: Effects on Learning and Retention for Students of Different Ability Levels

Nancy Russell Hamilton

Grant Number Title VII: 7-48-7670-204
National Defense Education Act of 1958
Leslie J. Briggs, Principal Investigator

Final Report Experiment I

June 1965

The Research Reported Herein Was
Supported by a Grant from the
U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education

Abstract

An autoinstructional program on the subject of Communism with reading-only response versus overt response was administered one hour each day for three weeks, to six classes of eighth grade students. Review programs were also given, with either reading-only response or overt response, and under conditions of either spaced or massed practice. Different kinds of measuring instruments were employed--each with the intent to assess learning from some separate aspect of the subject matter: General, Chronological, or Conceptual. Measurements of *Ss*' knowledge were made for pre-acquisition, post-acquisition, post-review, and for delayed-retention after a summer vacation. All analyses were made in terms of three levels of student ability. For gain and retention of purely chronological information, results showed that overt responding on review materials was of significantly greater value than reading-only responding-as might be expected with this type of subject matter. However, on the general materials high ability students profited from reading-only response conditions as much as, if not more than from overt response conditions. For middle-ability students and general materials, a significant favorable effect was found from having a *different* mode of response on review than was employed in initial instruction. Overall, lower ability students learned much less than higher ability students, and in the final analysis, low ability students appeared to have retained little or nothing. No clear value was demonstrated either for or against the particular schedules of practice used in this experiment during review. No effect other than that owing to *Ss*' ability was found with the test of conceptual knowledge. Overt response versions of instructional materials took substantially more time than reading-only response versions.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to Dr. Dennis J. Doolin of the Department of Political Science, Stanford University for expert consulting on the subject-matter of the auto-instructional program; to Dr. Edwin B. Parker of the Communication Department, Stanford University for reading this manuscript and providing many helpful suggestions as to content and structure; and to Mr. Marshall Sylvan and Mr. Carl Morris of the Stanford University Statistics Department for their very efficient treatment of a difficult data analysis problem.

The study described in this report was made possible through the cooperation of the members of the School Board and Mr. L. M. Larsen, Superintendent, Belmont School District; Mr. Claude Turner, Principal, Ralston Intermediate School; Dr. Rex Turner, Superintendent, Mr. E. S. Elsen, Assistant Superintendent, Sequoia Union High School District, and Mr. Ivan A. Griffin, Principal, Carlmont High School.

The authors are indebted to Mrs. Rosalie Garrison, Mrs. Ruth Holmquist, and Mrs. Janice Jensen of Ralston Intermediate School for administration of materials and for their many helpful suggestions during the course of the research.

Special thanks go to Mr. James Shearer, who made many of the graphic depictions, to Mrs. Shirley Bushnell and Mrs. Madalynne Chapman who provided very generous assistance during the more difficult and less rewarding phases of this research, and to Miss Sally Richerson and Miss Brenda Bordwell for preparation and typing of manuscripts.

Table of Contents
Experiment I

Page

Problem. - 1

Literature on Factors Which Influence Retention. - - - - - - - - - 1

  Loss of Knowledge through "Forgetting". - - - - - - - - - - - 1

  What Kind of Review Should be Provided? - - - - - - - - - - - 2

  Literature on Independent Variables Employed in this Study. - 3
    Response Mode. - 3
    Distribution of Practice. - - - - - - - - - - - - - - - 4

  Literature on the Influence of Ss Ability. - - - - - - - - - 5
    Studies of Ability Interactions with Persuasive Communications. - - - - - - - - - - - - - - 5
    Studies of Ability Interactions with Overt or Covert Responding. - - - - - - - - - - - - - 5
    Studies of Ability Interactions with Scheduling of Practice or Review. - - - - - - - - - - - 6
    Studies of Ability Interaction over Periods of No Treatment. - - - - - - - - - - - - - - 6

Measurement of Retention. - - - - - - - - - - - - - - - - - - - 6

  Different Kinds of Measuring Instruments. - - - - - - - - - - 6

  Recall versus Recognition. - - - - - - - - - - - - - - - - - 7

Summary of Hypotheses. - 7

  Main Effects. - 7

  Interactions with Level of Ability. - - - - - - - - - - - - - 8
    Independent Variables. - - - - - - - - - - - - - - - - - 8
    Changes over Time. - - - - - - - - - - - - - - - - - - - 8

Method. - 9

  Design. - 9

  Subjects and Assignment. - - - - - - - - - - - - - - - - - - 9

  Materials. - 9
    Programs for Initial Instruction. - - - - - - - - - - - - 9
    Review Programs. - 10
    Measuring Instruments. - - - - - - - - - - - - - - - - - 12

Table of Contents
Experiment I (continued)

Page
Procedure. - 12
    Pretesting and Initial Instruction. - - - - - - - - - - - 12
    Review. - 13
    Testing. - 13
    Multiple-Choice and Open-End Test Formats. - - - - - - 14

Results. - 14

Data. - 14

Analysis of the Data. - 14

Gain over the Entire Experiment. - - - - - - - - - - - - - - - 16

Effects at Delayed-Retention Testing. - - - - - - - - - - - 19

Effects of Review. - 27
    Short-Term Effects of Review. - - - - - - - - - - - - - - - 27
        The Long General Test. - - - - - - - - - - - - - - - - - 27
        The Chronology Test. - - - - - - - - - - - - - - - - - - 27
    Long-Term Effects of Review. - - - - - - - - - - - - - - - - 29
        The Long General Test. - - - - - - - - - - - - - - - - - 29
        The Chronology Test. - - - - - - - - - - - - - - - - - - 34
        Summary of the Comparison. - - - - - - - - - - - - - - 39

Retention Losses During Periods of No Treatment. - - - - - - 40

Further Examination of Results. - - - - - - - - - - - - - - - - 43
    Ability and Differential Gain. - - - - - - - - - - - - - - - 43
        The Long General Test. - - - - - - - - - - - - - - - - - 44
        The Chronology Test. - - - - - - - - - - - - - - - - - - 48
        The Concepts Test. - - - - - - - - - - - - - - - - - - - 48
    The Significant Text by Review by Ability Interactions. - - - - - - - - - - - - - - - - - 48

Other Analyses. - 52
    Time to Complete the Instructional Program. - - - - - - 52
    Teacher and Classroom Differences. - - - - - - - - - - - 53
    Ability and Different Measuring Instruments. - - - - - 53
    Multiple-Choice and Open-End Test Formats. - - - - - - 54
        Gain Scores and Test Format. - - - - - - - - - - - - - 54
        Reliability. - 54

Discussion and Conclusions. - - - - - - - - - - - - - - - - - - - 55

Overt versus Reading-Only Response Mode. - - - - - - - - - - - 55

Table of Contents
Experiment I (continued)

Page

Spaced versus Massed Practice. - 57

The Relationship between Instructional Materials and Criterion Tests. - - - - - - - - - - - - - - - - 58
The General Tests. - - - - - - - - - - - - - - - - - 58
The Chronology Test. - - - - - - - - - - - - - - - - 59
The Concepts Test. - - - - - - - - - - - - - - - - - 59

Efficiency. - 60

The Influence of Student Ability. - - - - - - - - - - - - 60

Long-Term Retention: Comparison of This Study with the Results of Other Studies. - - - - - - - - - 61

References. - 63

Appendix A. Experimental Population - - - - - - - - - - - - 66
B. Analysis of Variance Tables - - - - - - - - - - - 68
C. - 77
Samples from the Initial Instructional Texts
The Delayed-Retention Test

INCREASING LONG-TERM RETENTION OF KNOWLEDGE

Experiment I: Reading Response versus Overt Response Programs for Initial Instruction and Review: Effects on Learning and Retention for Students of Different Ability Levels

Problem

The purpose of the research reported here was to investigate techniques for developing and using instructional materials which would raise the initial level of knowledge, increase retention through review, and decrease time spent by the student in the instructional setting. Self-instructional programs were the vehicles for both initial instruction and for review. All instructional materials were programed, in most senses of the word, and instructional sequences designated as "reading-only" were the same as the sequences in which overt responding was requested, except that the answers were supplied.

The major experimental variables examined herein were: 1. instructional program and review program format, with regard especially to overt versus covert response mode; 2. scheduling of the review, in spaced versus massed sessions. Minor experimental variables were: 1. different measuring instruments to assess different kinds of learning (a chronological items test, a general items test, and a conceptual items test); 2. different test formats (multiple-choice versus open-end). All variables were examined with regard to ability of Ss.

Literature on Factors Which Influence Retention

Loss of Knowledge Through "Forgetting".

Several studies on the normal course of forgetting are reported in a text by Pressey and Robinson (1944). The text presents (page 543) a graphic representation of results of a number of early experiments on retention of school material (Bassett, 1928; Green, 1931; Johnson, 1924; Powers, 1925). The graph shows that over a period of seven months following the immediate post-acquisition test, a loss in measured knowledge of at least 12% was found for each of the population groups studied and for each of the subject-matters used. It was suggested that retention of material from a one-semester course (when advanced courses do not follow) is so poor as to cast doubt on the wisdom of offering such courses. An alternative to eliminating such courses from the curriculum is, of course, to improve the introductory course itself in a manner which would establish initial learning more firmly and thereby increase retention.

In most of such investigations, questions must be answered concerning basic objectives of school course-work, i.e., what is reasonable and useful for the student to retain. It was found (Pressey, et al., 1927), for instance, that simple usages in English composition such as the use of the apostrophe to show possession, taught from the third grade through the college freshman year, were still not known by college students, and only 73 percent of college freshmen could find the nouns in a sentence, while only 53 percent could locate the object of a verb in a simple sentence. It could be effectively argued that these students never "learned" these usages in the first place, and also, that the skills of locating the object of a verb or finding the nouns in a sentence are not skills essential to everyday use of the language. Possibly, long-term retention tests which make assessments of long-term behavioral objectives would provide a fairer basis for making what often seem to be value judgments about the effects of instruction.

What Kind of Review Should Be Provided?

Since a high level of initial learning and adequate review are both considered to be major factors in promoting retention, it was thought that use of autoinstructional programs for both purposes might result in significant improvements over the normal degree of long-term retention. Much research has been previously devoted to the use of autoinstruction for improving initial learning; with the present study, some applications have been made of autoinstructional techniques and materials for review learning as well.

There has been little discussion in the literature concerning what constitutes good or adequate review, although some positive effect can usually be demonstrated for a review procedure. It is possible that well-programed (in the broader sense) initial instruction might increase the level of initial learning, and/or establish more firmly the student's initial level of knowledge. Fewer review sessions might then be necessary to yield high levels of delayed retention. A number of investigations have been made which offer partial support for the foregoing premises. Naylor and Briggs (1961) concluded that there was a limit to the effectiveness of an increased number of trials provided during original training to facilitate retention of learned skills. Naylor, et al. (1962) concluded that absolute retention differences are primarily due to the degree of original skill acquisition. Brown, Naylor, and Briggs, (1963) investigated ways of maintaining a skill. They found that performance benefits which ordinarily were derived from review of the more difficult aspects of the task could not be demonstrated if training time was increased (more times per session) during the original learning situation. Bassett (1928) showed that degree of retention is a function of the level of initial learning, as might be expected; and evidence was obtained by Adams and Reynolds (1954) that the number of trials during initial practice was a factor having a crucial effect on long-term retention. (These findings were of particular interest and application in interpreting the results from the present study.)

An early study by Jones (1923) showed that retention levels ordinarily dropped to 30% of the amount initially learned during the first two months after acquisition, but the administration of a five-minute test following a lecture improved retention substantially for both "thought questions" and "fact questions". Sones and Stroud (1940) found that re-reading material was a better review method than taking a test several weeks after learning, while taking a test was better soon after learning. In the present experiment, re-reading nearly a month's-worth of material as review would have been quite impractical. However, guided partly by the results of the above studies, short review tests (with confirmation provided) were administered at the end of each section in the initial instructional programs, in addition to the formal review programs which were given at various intervals following the acquisition phase of the experiment.

Literature on Independent Variables Employed in this Study.

Response Mode.

A great many investigations of the form-of-response issue have been made in programed instruction research. Some investigators have found overt responding to be superior to nonovert modes of responding (Holland, 1960; Williams, 1963; Krumboltz and Weisman, 1962), while others have not found any differences in criterion performance between overt and nonovert modes (Alter and Silverman, 1962; Evans, Glaser, and Homme, 1962; Keislar and McNeil, 1962; Lambert, Miller and Wiley, 1962; Stolurow and Walker, 1962). Some of the literature dealing with the overt versus covert response issue led the author to believe that different effects might appear from overt or covert response, depending, among other things, upon the point in time at which assessments of these effects were made. In one study (Goldbeck and Campbell, 1962), it was found that on a 10-week retention test, S̲s who had used a reading version of a program retained more than S̲s who had used an overt-response version of the program (see recent reviews by Briggs and Angell, 1964; and Briggs and Hamilton, 1964). A number of explanations have been advanced to aid in the interpretation of these apparently contradictory findings. For instance, programs with different subject-matters may very easily produce different results in terms of benefits from overt response (see Eigen, 1961; Briggs and Hamilton, 1964).

A more theoretical explanation for conflicting response-mode findings was suggested as an outgrowth of the process used in contructing the instructional instruments for the present study: The instructional program was designed to inform eighth-grade students of some history and background of the Communist philosophy and political movement, as a basis for providing a clearer understanding of why the American system of government is preferable. With this type of subject matter it was very difficult to ascertain which kinds of responses would be "best" for overt practice. In this program, various procedures for obtaining overt response were employed: fill-in blanks, short answers, multiple-choice questions, matching, etc. However, many of these overt response procedures could have been accomplished differently and still have allowed the program to be classified as a program which required overt response. Contradictory evidence found in different investigations of response mode should, then, be examined partly with reference in each case to how arbitrarily different the decisions might have been as to which portions of the learning material were appropriate for overt practice, as well as to how arbitrarily different the mechanisms for producing overt response might have been, from those mechanisms which actually were employed in the experimental program. An earlier discussion (which was ancestral to the above) may be found in Lumsdaine (1961) p. 5-8.

With programing style held constant between the two versions of instructional materials, except for the overt response provisions, it was felt that the reading version would have an effectiveness that might not otherwise be expected in a reading-only text. Also, in considering the response-mode variable for the present experiment it was felt that the reading versions of instructional material would certainly take less time to complete than the overt response versions. (Investigations by Evans, Glaser, and Homme, 1959;

Shettel, Clapp, and Klaus, 1963; and Williams, 1963 found that reading versions of programs took less time to complete than did overt response versions.) Keeping in mind the upgrading of the instructional materials through the programing process as well as the economy of time which should result from the reading-only versions of programs, it was also expected that the reading versions might well be more efficient (amount learned or gained, taking time spent in acquisition into account) than overt response versions.

Discounting for the moment any interactions which might be hypothesized, as well as the conflicting response mode findings, the superior experimental treatments in terms of actual amount learned or gained should be those involving overt responding in either text, review, or both, and spaced rather than massed review sessions. In spite of the problems with the overt-covert response issue, a favorable effect for overt responding during review should be demonstrable at least for paired-associate types of materials (such as the chronology materials used in this study), even if not for the more general learning materials. We believed this to be the case because, as the programs were prepared, there was very little problem deciding what would be relevant for the student to practice of the paired-associate kinds of materials such as those in the chronology, for instance; while such decisions for general material were much more difficult, and much more arbitary; and there seemed to be more room for error.

Distribution of Practice.

The spacing and the form of review periods have been studied by a number of investigators. Spitzer (1939) found that frequent early reviews, followed by less frequent later reviews, could maintain retention at a useful level. He also found less positive effect from later review when this was not preceded by earlier review. In the present experiment review periods were spaced closer together in the beginning and further and further apart throughout the review schedule. It is possible, however, that if autoinstructional methods increase initial level of knowledge, and/or if the mental ability of Ss should be indicated as a factor in interpreting differential influence from experimental treatments, the optimum schedule of review would differ from that found in Spitzer's study. Adams and Reynolds (1954) found that the number of trials during initial practice was more important than the conditions of practice distribution.

Underwood (1961) summarized ten years of experiments designed to determine "the range of conditions and materials within which distributed practice facilitated learning or retention." (p. 229) For materials of low meaningfulness, he found that the extent of interference between rehearsals exerted enormous influence on the benefits of distributed practice--the greatest benefit from distributed practice deriving from schedules where interference was low. However, Underwood's findings for meaningful materials were less definitive, and his evidence indicated that distributed practice is likely to be more beneficial for materials of lower rather than higher meaningfulness.

Literature on the Influence of Ss Ability.

In the present research, ability level of the experimental population was a major variable, around which most of the hypotheses were centered and most of the analyses of data were made. It was decided to treat level of ability as an analysis variable, pooling the six separate classroom replication groups and then making divisions into three separate levels of ability for analysis.

The rationale for hypotheses about differential effects of experimental treatments on different ability-level Ss, was derived from a number of previous investigations where such differential effects have been found. Findings from some of these investigations are reviewed below.

Studies of Ability Interactions with Persuasive Communications. One foregone--although fairly tacit--scholastic purpose of our instructional program was to intensify Ss' socially desirable opinions concerning the subject of Communism. Effects of a "one-sided" versus a "two-sided" presentation of an argument on a controversial subject were studied by Hovland, Lumsdaine, and Sheffield (1949, pp. 201-227). Their instructional instrument was of a kind distantly related to the one employed in the present experiment. Hovland and his co-workers found that most of their subject-population were more positively affected by the "two-sided" version of an argument. But, lower ability (low educational level) students, who had initially been favorable to the point of view of communication, were negatively affected by the two-sided version. The interpretation made of this was that probably the two-sided version simply aroused confusion by "answering" questions which had not even occurred to that group of lower ability Ss before.

Where persuasive materials were used in the present instructional instrument (Learning About Communism, Hamilton, et al., 1963) the arguments employed were essentially of the "two-sided" variety. For instance, some of the issues which make Communist philosophy appear benevolent were brought up, but these issues were discounted with strong evidence such as the failure of "pure" Communist philosophy to function in existing Communist political systems. Although the measuring instruments in our study dealt with facts, figures, and concepts, rather than with attitude and opinion change, it was thought possible that some such effect as observed in the Hovland, Lumsdaine, and Sheffield experiment might differentially influence the learning and retention of facts by students of different ability levels as they participated in the experiment.

Studies of Ability Interactions with Overt or Covert Responding. Hovland, Lumsdaine, and Sheffield (1949, pp. 228-246) found that overt responding in a simple paired-associates learning task had more advantage (over observing only) at the lower intelligence levels than it did at the higher intelligence levels. Wittrock (1963) also obtained evidence on the interaction of learning ability with response mode. He found that while children of average intelligence performed better when they had used an overt mode of responding during acquisition, rather than a reading mode, the reverse was true for children of above-average intelligence.

These findings were directly antecedent to hypotheses for our present study, which stated that (a) higher ability Ss would be able to learn proportionally more than lower ability Ss, from the treatments expected to be generally less effective (reading-only response mode and massed review sessions), and that (b) lower ability Ss would profit proportionally more than would higher ability Ss from the treatments hypothesized to be generally more effective (overt response mode and spaced review sessions).

Studies of Ability Interactions with Scheduling of Practice or Review. Not a great amount of literature exists which is pertinent to our needs on this topic, because much of the work that has been done employed learning materials and time scheduling different from our own. However, additional background support for the focus on interactions with ability may be found in a study by Madsen (1963) who obtained evidence of an interaction between intelligence and a scheduling-of-practice variable. With a paired-associate learning task it was found that distributed practice trials were more beneficial than massed trials for low ability Ss, but that this was not true for high ability Ss.

Studies of Ability Interactions over Periods of No Treatment. Alter (1963) found differential changes between testing points (lines connecting test points had different slopes) for Ss of different ability levels when some experimental treatment intervened. However, no differential changes among ability levels could be substantiated when only time intervened between testing points (line slopes were not discriminably different).

See Figure 12, page 42, which was constructed as a model for the following hypotheses. For the present study, a prediction was made that slopes of the lines connecting the testing points where experimental treatments had intervened should be positive, and that lines connecting testing times for higher ability students should have greater slope than lines connecting testing points for lower ability students. That is, the line slopes should be positive from pre-acquisition to any subsequent testing time and from post-acquisition to any subsequent testing time, but not from post-review to the delayed-retention testing time. This hypothesis was constructed specifically to deal both with immediate changes in behavior, and with those changes which should accrue owing to the cumulative effects on the different ability levels, of instructional procedures from both text and review. (High ability will learn more and will gain more than low ability.)

Concerning changes between the post-review test and the delayed-retention test, since only time would intervene between these two testing points, it was felt that lines connecting these points would have negative slope. The slope should be nearly equal for students of all abilities, but each line would have different placement, according to its post-review point of origin. (Low ability will not lose more than high ability.)

Measurement of Retention

Different Kinds of Measuring Instruments.

Luh (1922) showed that retention level appeared to be different when different methods of measurement were used. The results of experiments such as Luh's have usually been interpreted in terms of the "sensitiveness" of various ways of measuring retention (McGeoch, 1946, p. 341), and little practical application of such findings has been attempted. The problem of generalizing findings from studies, when each research was conducted with different instructional materials and when results for each were obtained with different measuring instruments, has been discussed by several writers (see a review of literature by Briggs and Hamilton, 1964).

In the present research, for the purpose of demonstrating the different effects on learning which might be shown by using separate measuring instruments for different kinds of subject matter, three kinds of criterion tests were employed: (1) two tests of general information, (2) a test of chronological facts, and (3) a test of concept learning. It was expected that these different facets of the Ss' learning (general, chronological, and conceptual) would be highlighted by each of the different tests. For instance, with more or less one to one associates as with chronological materials in this study it is very clear which are the relevant terms for overt practice. Owing to this reasoning it was felt that learning of chronological paired-associates would show greater benefit from overt practice, than would learning of the more difficult-to-program general and conceptual materials.

Recall versus Recognition.

Test forms employing recognition have been considered by some investigators to be more sensitive measures of retention than forms employing recall (McGeogh, 1946). Landauer (1962), using listed paired-associates as learning materials, found that higher scores were obtained on multiple-choice forms of a test than were obtained on parallel questions in open-end form. These scores were enough higher that they could not be attributed solely to a chance or guessing factor. However, because of extreme differences in the learning materials and in the research procedures between the present research and Landauer's, specific predictions were not made that m-c test forms would produce differentially higher test scores than would o-e forms (i.e., higher than expected chance factors could produce). For the present experiment it was simply hypothesized that scores on multiple-choice forms of criterion tests would be higher in general than scores on parallel open-end forms of criterion tests.

Summary of Hypotheses

Main Effects.

If main effects occurred without regard to the ability level of the students, it was expected that overt response should produce higher criterion test scores than reading-only response on initial instruction, review instruction, and both initial instruction and review in combination. In connection with the response mode variable a sub-hypothesis was made that a favorable effect from overt responding during review should be demonstrable at least with paired-associates types of materials, such as those in the chronologies.

It was expected that spaced review sessions should produce greater learning than the massed sessions, and this effect should be more pronounced for paired-associates kinds of materials.

It was expected that higher ability Ss would have generally higher criterion test scores than lower ability Ss.

It was expected that reading-only conditions would take substantially less time than overt response conditions.

The reading-only versions of instruction were programed (using student tryout and revision, etc.) because the reading-only versions were simply the overt response versions with the responses fitted in as text. The assumption that reading-only forms of the instructional instruments would--through virtue of the programming process--be more effective than an ordinary textbook was not tested in this experiment. However, this assumption, combined with the expectation that reading-only conditions would take substantially less time, was the basis for the prediction that reading-only conditions should be more efficient (score on criterion test divided by instructional time) than overt responding conditions.

It was expected that multiple-choice test forms would allow higher criterion test scores than would the parallel open-end forms.

Interactions with Level of Ability.

Independent Variables.

It was predicted that higher ability Ss would be able to learn proportionally more than would lower ability Ss, from the treatments expected to be generally less effective (reading-only response mode and massed review sessions). It was also predicted that lower ability Ss would profit proportionally more than would higher ability Ss from the treatments hypothesized to be generally more effective (overt response mode and spaced review sessions).

Changes over Time.

It was predicted (Figure 12, p. 42) that where experimental treatments (initial instruction and review) had intervened between testing points, the slopes of lines connecting the testing points would be positive; in these cases it was predicted that lines connecting testing times for higher ability students should have greater slope than lines connecting testing points for lower ability students. (These lines would not be parallel.)

It was predicted that lines connecting testing points where only time intervened would have negative and nearly equal slope for all abilities, with the different starting and ending points reflecting only the absolute difference between the ability levels at the time of post-review testing. (These lines would be parallel.)

Method

Design.

The paradigm for the 2 x 2 x 2 design of the experiment is shown in Figure 1, below.

Figure 1

Experimental Design

Review Instruction		Initial Instruction	
		Overt Response Program	Reading-Only Program
Overt Response Program	Spaced Review	1	2
	Massed Review	3	4
Reading-Only Program	Spaced Review	5	6
	Massed Review	7	8

Subjects and Assignment.

Six eighth-grade Social Studies classes participated in the study. (N total at pre-acquisition was 201, including 23 no-review students.) A complete replication of the experiment--involving eight experimental treatments--was made in each class. Three Ss in each class were assigned at random to each of the eight treatment cells in the experiment. After the randomization and assignment, the Ss who remained were designated to be a sort of "control group" (receiving no review treatment) for the study. There were from one to seven Ss of this type in each class, depending on the total number of students in the class, making 21 no review Ss in all. Half of these control Ss were assigned at random to each of the two initial acquisition text versions. If at any time before the beginning of the first review treatment a cell in the experiment lost an S, one of the Ss from this control group was assigned to the depleted cell to complete the experiment. (The total N at delayed-retention testing was 127, excluding 21 no-review students. This attrition rate, occurred over the summer through some students leaving the area, some students going to other high schools, etc.)

Materials.

Programs for Initial Instruction.

The instructional instrument was a self-teaching program on Communism (Hamilton, et al., 1963), which was prepared by the research staff especially for this project. Curriculum guides were provided by the participating school district. An expert consultant on the subject-matter of the text verified the accuracy of the content and advised on matters of organization and emphasis. The basic program, with overt response, was tried out with individual students and revised several times until it yielded satisfactory performance by a group of six students on a short criterion test.

The format of the program was linear (progressing from one step to the next without any detours or recycling). The student was required to make a wide variety of active responses in the overt response version, including filling-in-the-blanks, constructing the entire answer to a question, choosing the correct answer to a question, and matching names, dates, places, and figures. Confirmation was provided for each response. The reading version of the program presented the same material, in the same page format, page for page, but each occasion for overt response was replaced with the material which had been used in the overt versions as confirmation. Each version contained 377 pages and comprised three to twelve hours of programed material, depending upon the rate of individual student progress. Samples from each program version may be found in Appendix C.

Review Programs.

The review programs were constructed on the assumption that exposure to the material during initial instruction provided sufficient background to the student to enable him to answer questions about it. It should be emphasized that the programing of these review materials was not the same as it would have been for initial instructional information. Conceptual and factual materials were prepared in a question and answer format. The procedure for developing review programs was as follows:

(1) Each chapter of the program was searched for conceptualizations. These were phrased in question form. (See the sample item "What is an ideology?", below.)

(2) Each page of the program was concurrently searched for informational and factual material. (See the sample item "An ideology is a set of...", below.)

(3) The informational and factual material was organized under the conceptual question headings.

(4) The project staff and consultants then edited these materials--eliminating redundancies, reorganizing, checking for structural flaws, etc. (All of these efforts were made as an attempt to insure complete coverage of materials for review of the review material.)

The overt response version of the review contained items much like criterion frames in the initial program, and the student filled in the blanks with material that he was already supposed to have learned. Confirmation was supplied for an entire page of questions and answers on the reverse of the page. Reading-Only Ss simply read the question and answer sequence, filled-in as an outline. A sample review item in the Overt-Response format is:

"What is an Ideology?

1. An ideology is a set of strong__________ or __________ about something.

2. Communism is an__________."

On the reverse of the page, this confirmation appears:

"What is an ideology?

1. convictions
   beliefs
2. ideology"

In the Reading versions of these materials, the same item appeared as follows:

"What is an ideology?

1. An ideology is a set of strong convictions or beliefs about something.
2. Communism is an ideology."

In addition to the general review program (which was divided into three sections for administration purposes), special materials were prepared for review of chronological information. This chronology consisted of significant dates, names, and events from Russian and Chinese history. All of the material in this (as in the other review programs) was present in the original instructional text; however, unlike the materials which were covered in the general reviews, no concerted attempt was made in the original instructional text to have students memorize specific chronological items. So, for all practical purposes, both the Overt and the Reading response groups had reading-only versions of chronological materials in the original instructional text. The Russian chronology contained forty-one items and the Chinese chronology contained twenty-eight items. For the students receiving the reading version of the review program, twenty-one of the more important items in the listed Russian and Chinese chronologies were marked with an asterisk (*). The listed chronology for students receiving overt response review did not include the items marked with asterisks; rather, the information contained in the marked items was presented as a memorization (repeated practice) program, with overt response and confirmation.

The chronology itself was straightforward, factual and nonnarrative. The items were of an essentially one-to-one associates nature (e.g., "Karl Marx was born in Germany in 1818;" "The first Five-Year Plan was inaugurated by Stalin in 1928"). The Overt-Response groups, instead of merely being told to memorize marked items, were given programs which provided practice three different times (with the other items interspersed) on each item. Each of the three times, a different member of the set was left for recall (i.e., "4. Karl Marx was born in_________in 1818;" "5. _______inaugurated the first Five-Year Plan in 1928;" ..."9. Karl Marx was born in Germany in____(date);"..."14. The first_____-_____Plan was inaugurated by Stalin in 1928;"..."21. ____ ______ was born in Germany in 1818".)

All the students were told that while they might be tested on any item in the entire chronology, the marked information (*) for Reading Ss or the programed information for Overt-Responding Ss was of particular importance and should be memorized. However, information to the Ss notwithstanding, the Chronology Test measured their knowledge only on items which had been marked (*) or programed. No test items based on marked chronology information were allowed in the sampling pool for the General tests. However, a few of the other chronological points were on the General tests as a result of the random sampling from the general item pool.

Measuring Instruments.

Four tests were constructed: two tests for general knowledge, one test for chronological knowledge, and one test for conceptual knowledge. For the General tests, a population of 350 general test items was developed. These items were designed to assess Ss' knowledge of a reasonably complete amount of information from the instructional instrument--such as the classroom teacher might expect. To obtain these items, four staff members went through the entire program, and each made as complete an enumeration as he was able of informational facts and concepts in the program. These were made into test items, examined again, and the extremely easy, or "give-away" items were eliminated or modified. From this set, two separate General tests were sampled:

(1) A Short General Test of 20 items, requiring 27 responses.

(2) A Long General Test of 40 items, requiring 57 responses.

The Short General Test was constructed by a random sampling of items from the general item pool, then those items were returned to the sampling pool. None of these items happened to be in the sampling for the Long General Test, however, so there were no items in common between the two general tests. In addition, two tests were constructed from material which, from the standpoint of reasonable pedagogy would be in many instances, too detailed (chronological) or too difficult (conceptual) for most eighth-grade students. These instruments were used to provide more sensitive tests for some of the hypotheses.

(3) The Chronology Test consisted of 17 items which called for 46 responses. These items measured in detail all the information included in the marked (*) or programed portions of the Chronological Review program.

(4) The Concepts Test consisted of 10 items which were allotted 46 score-points altogether. These items were chosen specifically to measure retention of conceptual material and more difficult factual material.

This procedure produced four measuring instruments. Each instrument was intended to be fairly high in difficulty (no give-away items were allowed in the sampling pool). The most difficult was the Concepts Test, judged by pre-experimental trials and from the results measured with this instrument. See measuring instruments in Appendix C.

Procedure.

Pretesting and Initial Instruction.

Immediately before initial instruction was begun with either the Overt Response or the Reading version of Learning About Communism, the Short General Test was administered to all Ss as a pretest. Then, under the supervision of the regular classroom teacher, Ss studied the program for about 40 minutes each school day. It was possible to determine the time spent in initial instruction, because each S kept a running record of the time he spent on each chapter of the program by stamping the booklet in an electronic time recording device as he started and stopped work. As soon as a student finished the program he took the Long General Test and the Chronology Test (the Concepts Test was not prepared before the post-review testing). The pretesting, administration of the initial learning program, and the immediate retention testing took approximately two and one-half weeks in each classroom.

Review.

The Chronology and each of the three sections of the general review program were covered twice during the review phase by all treatment groups. Review lessons were presented in either Overt Response or Reading-Only form, and under either massed or spaced conditions. The review pattern, indicating two repetitions for all Ss, looked something like this:

Spaced: x x x x x x x x
Massed: xxxx xxxx

The Ss in the Massed Practice treatment had four review sessions, the first two beginning 20 days after all students had completed initial instruction, and the second two beginning 46 days after the completion of the first session. These sessions took place for two days running, with students working for 100 minutes without break on each of the days. In the Spaced Practice conditions, Ss had eight review sessions, the first beginning four days after all students had completed initial instruction, and the others following--each was supposed to occur at a longer time interval than the interval preceeding it. (During the first complete review, lessons were presented at intervals of 8, 10, and 11 days apart. During the second complete review, however, lessons were presented at intervals of 21, 15, and 16 days--largely due to difficulties with school scheduling.) The review sessons in the Spaced Review treatments were, about 50 minutes long.

Testing.

Figure 2 shows the times at which the various tests were administered. The pre-acquisition test and the post-review tests were administered to class groups. The post-acquisition tests were administered to each S individually, as he completed the learning program.[1] The delayed-retention tests were administered to all Ss assembled together in a single, large group.

Figure 2
Times During the Experiment at Which Various Tests were Administered

Test	Time of Testing						
	Pre-Acquisition	3 Weeks	Post-Acquisition	15 Weeks	Post-Review	15 Weeks	Delayed-Retention
Short General	X	Acquisition Time		Review Time	X		X
Long General			X		X		X
Chronology			X		X		X
Conceptual					X		X

---

[1] We decided it would be better to have the immediate-retention testing be truly immediate, since some students would have had as long as a week and a half delay if we had waited until everyone was finished. This procedure risked that students who had completed the test would recall items and discuss them with companions who were still working. Teachers attempted to monitor this in class and did not report any great student interest in that kind of "cheating". We do not know, of course, what effects such prompting of one student by another might have had and can only hope that it was minimal.

Multiple-choice and Open-end Test Formats. As mentioned previously, to provide measures of recognition and recall, *S*s were further randomly divided into groups receiving as a pretest either a multiple-choice or an open-end format of the Short General Test. *S*s also received either m-c or o-e forms of the Long General Test and the Chronology Test at the post-acquisition testing. Assignment of *S*s to the two formats for the post-acquisition testing was the reverse of the assignment of *S*s to the formats for the pretesting. A single, combined form was administered to all *S*s at the time of delayed retention testing. Half of the items were selected at random from the multiple-choice format; the remaining half of the items were, therefore, in open-end format. This procedure allowed a delayed retention test of both recall and recognition, as well as providing a group of items comparable for every *S*.

Results

Data.

Program responses and time scores for program completion were used in addition to scores obtained from items answered correctly on criterion tests (raw scores, percent scores, and difference (gain or loss) scores).

Analysis of the Data.

Analysis was made of the entire experiment overall (six replications as one), among all classifications, and pooled among the three ability levels. At the time of the long-delayed retention testing, High Ability level contained the 41 *S*s who ranged from 92-100 percentile in total score on the *School and College Ability Test* (mean: 95); Medium Ability level contained 35 *S*s who ranked from 71-90 percentile on the SCAT (mean: 80); and Low Ability level contained 43 *S*s who ranged from 3-65 on the SCAT (mean: 40). To illustrate why it was appropriate to pool *S*s across classes and analyze by levels of ability, see Appendix A, where these distributions of ability level may be compared with the distributions of *S*s by classrooms.

Analysis of variance was employed; however, a disproportionality of observations among cells had resulted by the time of delayed-retention testing. Because the retention data and gain-loss data in general were considered of greatest importance to this experiment, a method for multiple classification analysis of variance with disproportionate N's was required. A full description of the method will be provided upon request. In general, an independent estimate of error variance is computed, using each observation in the entire experiment. After this estimate of error is obtained, the analysis is continued using only cell means, means of means, and a grand mean; no further reference is ever made to the single scores. There are a great many restrictions as to where this method might be sensible to apply, but if this method were used with data where there were *equal* N's in experimental cells, the results would be *exactly* equivalent to those obtained with an ordinary computational method. (In the other present instance, the computation with means provides a better estimate than other computational methods, of the values which would be obtained if there were equal (or proportional) N's in the cells.)

When a reliable F-ratio was obtained by this analysis, a significance of difference between treatments was assessed by a generalization of the critical difference method (Lindquist, 1953):

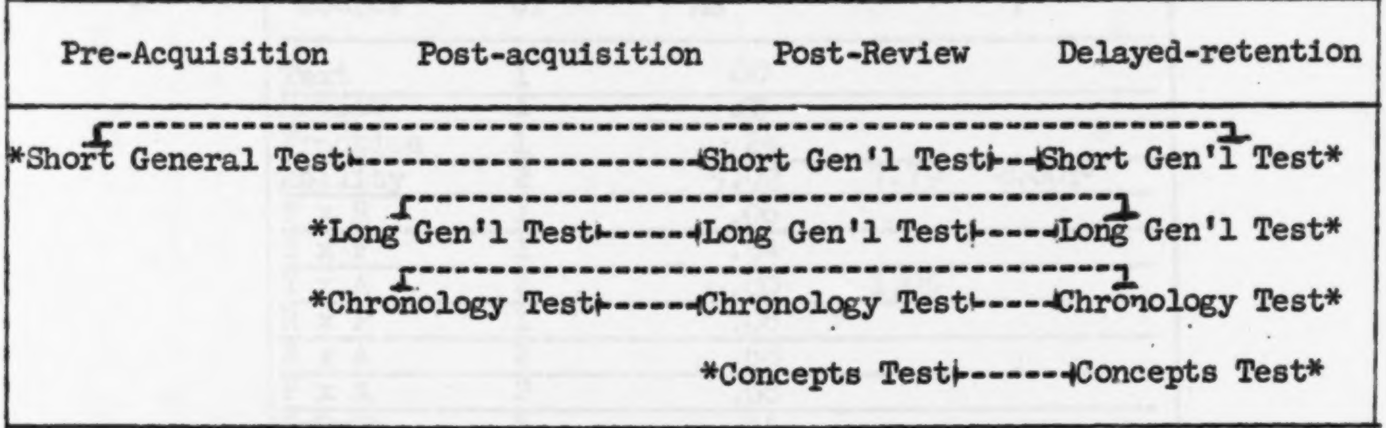
$$d = (t)\sqrt{\frac{2ms_{error}}{\text{mean cell-frequency}}}$$

The "t" value selected here and the generally accepted level of significance for all analyses in the experiment was for a probability of five percent. Analyses of variance are all four-way (2x2x2x3) unless otherwise noted.

Diagram 1, below, shows the comparisons made among treatments by the four measuring instruments.

Diagram 1

Analyses made of Criterion Tests.


Pre-Acquisition
Post-acquisition
Post-Review
Delayed-retention
*Short General Test
Short Gen'l Test
Short Gen'l Test*
*Long Gen'l Test
Long Gen'l Test
Long Gen'l Test*
*Chronology Test
Chronology Test
Chronology Test*
*Concepts Test
Concepts Test*

(*) indicates that a raw scores analysis was made; a connecting line. (-----) shows that a difference-scores analysis was made.

The following abbreviations have been used in discussing these analyses:

T: Text, Overt Response or Reading-Only response mode

R: Review, Overt Response or Reading-Only response mode

A: Ability; High, Medium or Low ability level

HA: High Ability, referring to the ability level of the Ss

MA: Medium Ability

LA: Low Ability

OT: Overt Response Mode on the Text

RT: Reading-Only Response Mode on the Text

OR: Overt Response Mode on the Review

RR: Reading-Only Response Mode on the Review

SP: Spaced Practice on Review

MP: Massed Practice on Review

Gain over the Entire Experiment.

Effects of the experiment overall were measured by gain scores on the Short General Test, from pre-acquisition to delayed-retention. Table 1, below (also Table I in Appendix B) reflects an effect owing to Ability level ($p < .001$), but no main effects owing to the independent variables. There was, however, a significant interaction ($p < .025$) of response mode on initial text, with response mode on review, with level of ability (T x R x A).

Table 1

Analysis of Variance indicating effects measured by the Short General Test, from pre-acquisition to delayed-retention.

Source	df	MS	F	P
Text	1	.00		
Review	1	.00		
Practice	1	.24		
Ability	2	5.92	7.79	<.001*
T x R	1	.06		
T x P	1	.54		
T x A	2	.80	1.05	
R x P	1	.06		
R x A	2	.68		
P x A	2	.68		
TxRxP	1	.15		
TxRxA	2	2.99	3.93	<.025*
TxPxA	2	.64		
RxPxA	2	.13		
TxRxPxA	2	1.94	2.55	<.10
ERROR	95	.76		

* indicates a difference accepted as reliable

Critical differences for this interaction are presented in Table 2. Critical difference for $p < .05$ is 1.03. An (*) indicates a difference which has reached or exceeded this critical difference.

In figure 3 is a rank-order expansion of the same T x R x A interaction shown in Table 2.

Table 2

Critical Differences, Text x Review x Ability Interaction ($p < .025$), Short General Test: Gain from Pre-Acquisition to Delayed-Retention

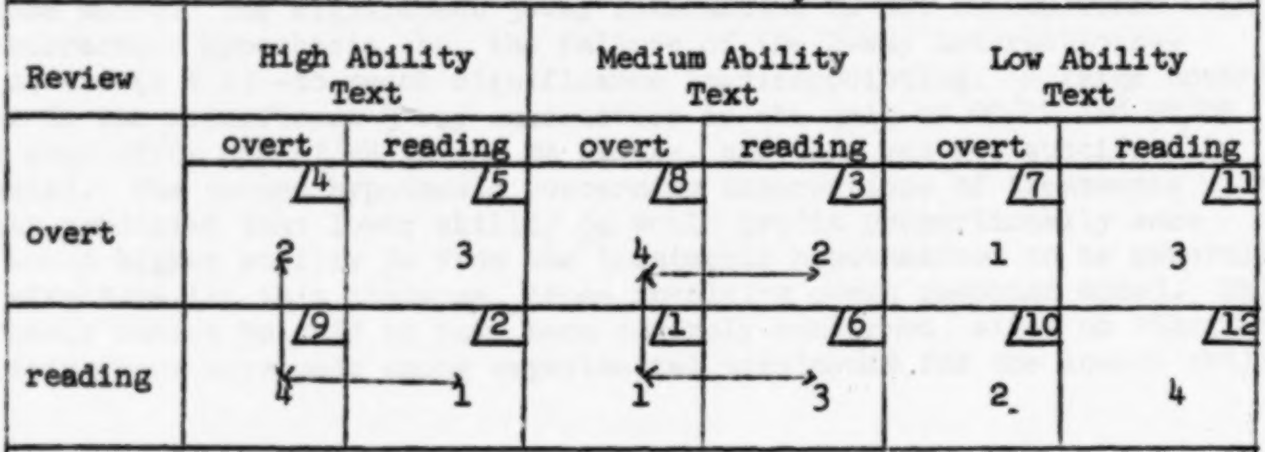
Ability				High				Medium				Low			
	Text			OT		RT		OT		RT		OT		RT	
		Review		OR	RR	OR	RR	OR	RR	OR	RR	OR	RR	OR	RR
			Means	3.00	1.84	2.74	3.35	1.96	4.14	3.34	2.60	2.04	1.28	1.20	1.07
High	OT	OR	3.00		1.16*	.26	-.35	1.04*	-1.14*	-.34	.40	.96	1.72*	1.80*	1.93*
		RR	1.84			-.90	-1.51*	-.12	-2.30*	-1.50*	-.76	-.20	.56	.64	.77
	RT	OR	2.74				-.61	.78	-1.40*	-.60	.14	.70	1.46*	1.54*	1.67*
		RR	3.35					1.39*	-.79	.01	.75	1.31*	2.07*	2.15*	2.28*
Medium	OT	OR	1.96						-2.18*	-1.38*	-.64	-.08	.68	.76	.89
		RR	4.14							.80	1.54*	2.10*	2.86*	2.94*	3.07*
	RT	OR	3.34								.74	1.30*	2.06*	2.14*	2.27*
		RR	2.60									.56	1.32*	1.40*	1.53*
Low	OT	OR	2.04										.76	.84	.97
		RR	1.28											.08	.13
	RT	OR	1.20												
		RR	1.07												

*$p < .05$

Figure 3

Rank Orders: Short General Test, T x R x A Interaction, Gain from Pre-Acquisition to Delayed-Retention (p < .025)

Treatments for Each Ability Level*



Review	High Ability Text		Medium Ability Text		Low Ability Text	
	overt	reading	overt	reading	overt	reading
overt	4 2	5 3	8 4	3 2	7 1	11 3
reading	9 4	2 1	1 1	6 3	10 2	12 4

Arrows indicate significant differences between treatments within each ability level. (⟷) Numbers centered in large boxes reflect the rank order of the treatments within each ability level. The number enclosed in the upper-right corner of each box reflects the position of that cell over the entire interaction. A rank of 1 indicates the best treatment.

It appears from examining this interaction that both the "Reading-Only Text/Reading-Only Review" (RT/RR) and the "Overt-Response Text/Overt-Response Review" (OT/OR) treatments were quite effective for high ability Ss. This is an especially desirable finding; it indicates that for acquisition of general information from this kind of subject-matter, high ability students may be provided with well-organized reading material, with no time-consuming overt response necessary and without fear of diminishing effectiveness. On the other hand should it be necessary for logistics reasons to administer overt response programs to high ability students, these results indicate that there would probably be no loss of effectiveness (although probably a longer time would be required for overt responding).

For medium ability Ss the two mixed treatments (i.e., Reading Text/Overt Review and Overt Text/Reading Review) were more effective than either Reading Text/Reading Review or Overt Text/Overt Review. The best condition for medium ability (OT/OR showed absolute gain significantly greater than most other conditions in the entire analysis, excepting only High Ability/RT/RR (ranked 2 overall) and Medium Ability/RT/OR (ranked 3 overall), from which MA/OT/RR (ranked 1 overall) was not reliably different.

For low ability Ss the four treatment conditions were not discriminably different from one another. Also, LA/RT/RR, which was the least effective treatment, in terms of size of mean, for low ability Ss showed absolute gain significantly less than for any other conditions in other ability levels, excepting only MA/OT/OR (ranked 8 overall) and HA/OT/RR (ranked 9 overall), from which LA/RT/RR (ranked 12 overall) was not reliably different.

The reader may recall the hypotheses concerning interaction of experimental treatments with ability level of Subjects. The first of these hypotheses predicted that higher ability Ss would be able to learn proportionally more than would lower ability Ss, from the treatments hypothesized to be generally less effective (in this instance, those involving reading-only response mode). The significant 3-way interaction is not inconsistent with the interaction hypothesis, but the failure of the 2-way interactions--(T x A) and (R x A)--to reach significance is disappointing. A large contribution to the significant 3-way interaction is the gain of OT/RR and RT/OR cells over OT/OR and RT/RR in the MA groups, and this was not specifically predicted. The second hypothesis concerning interactions of treatments with ability predicted that lower ability Ss would profit proportionally more than would higher ability Ss from the treatments hypothesized to be generally more effective (in this instance, those involving overt response mode). This hypothesis cannot be said to have been entirely confirmed, since no significant discriminations were made among experimental treatments for the lowest ability Ss.

Effects at Delayed-Retention Testing.

Long-term effects from experimental treatments were shown by the analysis of percent correct scores on the Long General Test.[2] The results of the analysis of variance are summarized in Table 3, below, and Table II in Appendix B.

---

[2] Analyses of percent correct scores at delayed-retention testing on tests other than the Long General Test, showed no significant effects other than that owing to ability level of the subjects. This single effect was reliable ($p < .001$) for all delayed-retention analyses. See the analysis of variance tables III, IV, V, VI, in Appendix B.

Table 3
Analysis of Variance indicating effects measured by the Long General Test, at delayed retention.

Source	df	MS	F	P
Text	1	2.61		
Review	1	32.57		
Practice	1	16.73		
Ability	2	2358.16	38.64	<.001*
T x R	1	90.56	1.41	
T x P	1	130.01	2.03	<.20
T x A	2	6.68		
R x P	1	38.41		
R x A	2	6.24		
P x A	2	135.44	2.12	<.20
TxRxP	1	717.77	11.21	<.005*
TxRxA	2	357.26	5.58	<.01*
TxPxA	2	38.36		
RxPxA	2	244.97	3.83	<.05*
TxRxPxA	2	.00		
ERROR	95	64.01		

*indicates differences accepted as reliable (p <.05)

This analysis indicated a significant main effect for Ability (p <.001). High ability students had the highest scores; and low ability students had the lowest. There were no significant main effects from other experimental variables. Three interaction effects were, however, found with this analysis: response mode on text, by response mode on review, by distribution of practice (T x R x P, p <.005); response mode on text, by response mode on review, by level of ability (T x R x A, p <.01); and response mode on review, by distribution of practice, by level of ability (R x P x A); p <.05). The critical difference between means for p=<.05 was 10.12 for each of these interactions.

Table 4 presents means and critical differences for the T x R x P interaction, and Figure 4 represents the interaction by rank orders.

Table 4

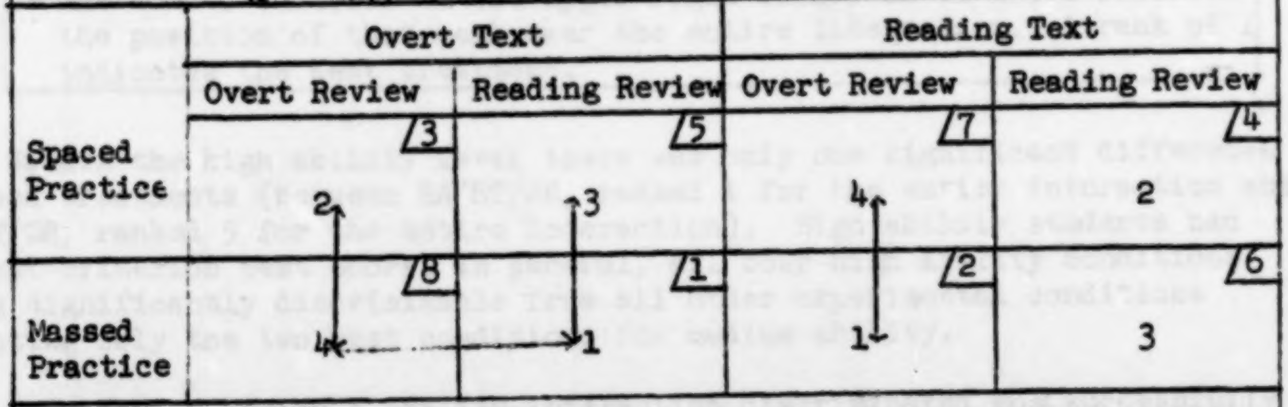
Critical Differences: Text x Review x Practice Interaction, ($p<.005$), Long General Test, Delayed-Retention, percent score correct.

Text	Review	Practice	Overt Text				Reading Text			
			Overt Review		Reading Review		Overt Review		Reading Rev	
			Spaced	Massed	Spaced	Massed	Spaced	Massed	Spaced	Massed
		Means	52.40	35.95	45.15	55.63	40.04	54.77	46.89	44.80
OT	OR	S 52.40		16.45*	7.25	-3.23	12.36*	-2.37	5.51	7.60
		M 35.95			-9.20	-19.68*	-4.09	-18.82*	-10.94*	-8.85
	RR	S 45.15				-10.48*	5.11	-9.62	-1.74	.35
		M 55.63					15.59*	.86	8.74	10.83*
RT	OR	S 40.04						-14.73*	-6.85	-4.76
		M 54.77							7.85	14.73
	RR	S 46.89								2.09
		M 44.80								

*P <.05

Figure 4

Rank Orders: Organized by Text Condition: Text x Review x Practice Interaction ($p<.005$), Long General Test, Delayed Retention.



	Overt Text		Reading Text	
	Overt Review	Reading Review	Overt Review	Reading Review
Spaced Practice	2 $\boxed{3}$	3 $\boxed{5}$	4 $\boxed{7}$	2 $\boxed{4}$
Massed Practice	4 $\boxed{8}$	1 $\boxed{1}$	1 $\boxed{2}$	3 $\boxed{6}$

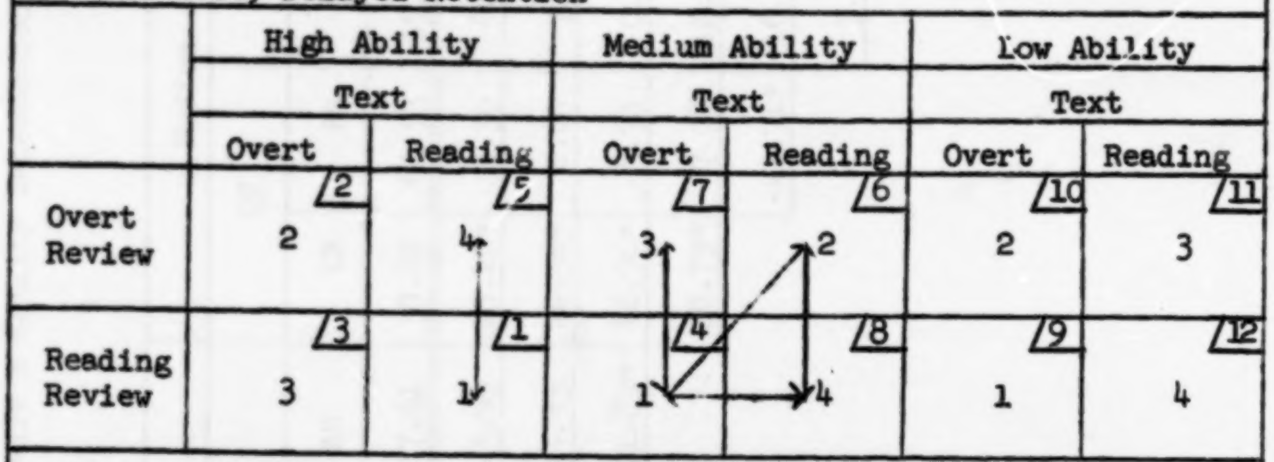
Arrows indicate significant differences between treatments within each ability level. (←——→) The numbers in the large boxes reflect the rank order of the treatments within each ability level. The number enclosed in the upper right corner of each box reflects the position of that cell over the entire interaction. A rank of 1 indicates the best treatment.

This interaction indicates that both the group having the highest criterion scores and the group having the lowest criterion scores had overt response on the text and massed practice (OT/RR/MP, ranked 1 overall and OT/OR/MP, ranked 8 overall). The crucial factor between conditions 1 and 8, then, appears to be response mode on review, reading-only response producing better criterion test scores than overt response. The best treatment which involved spaced practice was OT/OR (ranked 3 overall). These findings cannot be satisfactorily used to either confirm or refute hypotheses about Practice, Text, and Review because they are not consistent between treatment conditions.

Table 5 contains means and critical differences, and Figure 5 contains rank orders for the significant interaction of response mode on text, by response mode on review, by level of ability (T x R x A).

Figure 5

Rank Orders: Text x Review X Ability Interaction, ($p<.01$) Long General Test, Delayed Retention



	High Ability		Medium Ability		Low Ability	
	Text		Text		Text	
	Overt	Reading	Overt	Reading	Overt	Reading
Overt Review	$\boxed{2}$ 2	$\boxed{5}$ 4	$\boxed{7}$ 3	$\boxed{6}$ 2	$\boxed{10}$ 2	$\boxed{11}$ 3
Reading Review	$\boxed{3}$ 3	$\boxed{1}$ 1	$\boxed{4}$ 1	$\boxed{8}$ 4	$\boxed{9}$ 1	$\boxed{12}$ 4

Arrows indicate significant differences between treatments within each ability level. (←——→) The numbers in the large boxes reflect the rank order of the treatments within each ability level. The number enclosed in the upper right corner of each box reflects the position of that cell over the entire interaction. A rank of 1 indicates the best treatment.

Within the high ability level there was only one significant difference between treatments (between HA/RT/RR, ranked 1 for the entire interaction and HA/RT/OR, ranked 5 for the entire interaction). High ability students had highest criterion test scores in general, all four high ability conditions being significantly discriminable from all other experimental conditions excepting only the two best conditions for medium ability.

This Text x Review x Ability interaction discriminated most successfully within the medium ability level; the "mixed" treatment conditions (MA/OT/RR, ranked 4 overall and MA/RT/OR, ranked 6 overall) had the higher mean percent correct scores on the delayed retention test. The treatment ranked 4 overall

Table 5

Means and Critical Differences: Text x Review x Ability Interaction ($p < .01$), Long General Test, Delayed-Retention Percent Correct Score.

Ability	Text	Review	Mean	High				Medium				Low			
				OT		RT		OT		RT		OT		RT	
				OR	RR	OR	RR	OR	RR	OR	RR	OR	RR	OR	RR
			Mean	63.92	63.18	58.78	67.61	38.82	60.53	55.49	38.64	29.77	27.46	27.95	31.30
High	OT	OR	63.92		0.74	5.14	-3.69	25.10*	3.39	8.43	25.28*	34.15*	36.46*	35.97*	32.62*
		RR	63.18			4.40	-4.43	24.36*	2.65	7.69	24.54*	33.41*	35.72*	35.23*	31.88*
	RT	OR	55.78				-11.83*	16.96*	-4.75	0.29	17.14*	26.01*	28.32*	27.83*	24.48*
		RR	67.61					28.79*	7.08	12.12*	28.97*	37.84*	40.15*	39.66*	36.31*
Medium	OT	OR	38.82						-21.71*	-16.67*	0.18	9.05	11.36*	10.87*	7.52
		RR	60.53							5.04	21.89*	30.76*	33.07*	32.58*	29.23*
	RT	OR	55.49								16.85*	25.72*	28.03*	27.54*	24.19*
		RR	38.64									8.87	11.18*	10.69*	7.34
Low	OT	OR	29.77										2.31	1.82	-1.53
		RR	27.46											-0.49	-3.84
	RT	OR	27.95												-3.35
		RR	31.30												

*$p < .05$

had a significantly higher mean than any other condition within the middle ability level. MA/RT/RR, ranked 8 overall, had a significantly lower mean than any other condition within the medium ability level, excepting MA/OT/OR, ranked 7 overall, from which MA/RT/RR was not discriminably different.

There were no significant treatment differences within the low ability level.

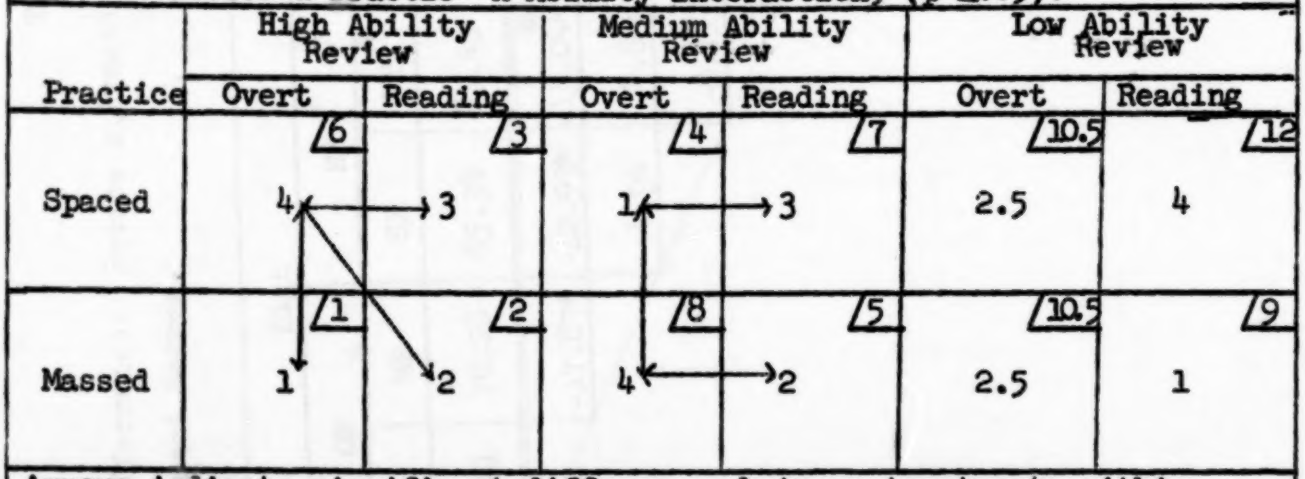
As might have been expected, high ability Ss had the highest overall mean, and low ability Ss had the lowest overall mean. The best treatment for high ability Ss was also the treatment taking least time (RT/RR), and this finding is consistent with the hypothesis stating that high ability Ss should be the subjects least detrimentally affected by reading-only response conditions.

This T x R x A interaction may be compared with the similar T x R x A interaction found with the Short General Test for overall gain scores from pre-acquisition to delayed-retention, discussed previously on pages 15-17. The results between these two analyses were quite similar, especially when it is noted that the two best conditions for each ability level were the same for both interactions.

The third significant three-way interaction in the analysis of delayed-retention scores on the Long General Test was for response mode on review, by distribution of practice on review, by level of ability (R x P x A; $p < .05$) Table 6 gives means and critical differences for this interaction, and Figure 6 gives rank orders of conditions within and among ability levels. In this

Figure 6

Rank Orders: Long General Test, Delayed Retention Raw Scores Review x Practice x Ability Interaction; ($p < .05$).



	High Ability Review		Medium Ability Review		Low Ability Review	
Practice	Overt	Reading	Overt	Reading	Overt	Reading
Spaced	6 4	3 3	4 1	7 3	10.5 2.5	12 4
Massed	1 1	2 2	8 4	5 2	10.5 2.5	9 1

Arrows indicate significant differences between treatments within each ability level. (←→) The numbers in the large boxes reflect the rank order of the treatments within each ability level. The number enclosed in the upper right corner of each box reflects the position of that cell over the entire interaction. A rank of 1 indicates the best treatment.

Table 6

Means and Critical Differences: Review x Practice x Ability Interaction ($p < .05$), Long General Test, Delayed-Retention Percent Correct Scores.

Ability				High				Medium				Low			
	Review			OR		RR		OR		RR		OR		RR	
		Practice		SP	MP	SP	MP	SP	MP	SP	MP	SP	MP	SP	MP
			Mean	52.41	70.28	65.34	65.45	57.38	36.93	46.59	52.58	28.86	28.86	26.14	32.62
High	OR	SP	52.41		-17.87*	-12.93*	-13.04*	-4.97	15.48*	5.82	-0.17	23.55*	23.55*	26.27*	19.79*
		MP	70.28			4.94	4.83	12.90*	33.35*	23.69*	17.70*	41.42*	41.42*	44.14*	37.66*
	RR	SP	65.34				-0.11	7.96	28.41*	18.75*	12.76*	36.48*	36.48*	39.20*	32.72*
		MP	65.45					8.07	28.52*	18.86*	12.87*	36.59*	36.59*	39.31*	32.83*
Medium	OR	SP	57.38						20.45*	10.79*	4.80	28.52*	28.52*	31.24*	24.76*
		MP	36.93							-9.66	-15.65	8.07	8.07	10.79	4.31
	RR	SP	46.59								-5.99	17.73*	17.73*	20.45*	13.97*
		MP	52.58									23.72*	23.72*	26.44*	19.96*
Low	OR	SP	28.86										0.00	2.72	-3.76
		MP	28.86											2.72	-3.76
	RR	SP	26.14												-6.48
		MP	32.62												

*$p < .05$

interaction significant differences were found between conditions within the high and medium ability groups, but again, not within the low ability group. The effects of the treatments were not consistent between ability groups. For high ability *S*s the Overt Review/Massed Practice (OR/MP) condition was superior (ranked 1 overall); while for medium ability *S*s the OR/MP condition was inferior (ranked 8 overall). Also, for high ability students, the Overt Review/Spaced Practice (OR/SP) condition was significantly poorer than all others; while for medium ability students OR/SP was significantly better than all other conditions save Reading Review/Massed Practice (RR/MP). In the RR/MP condition, however, we find the first consistency in the interaction; for all three ability levels RR/MP was either ranked 1 or ranked 2 within the ability level. This result may reflect the difficulty of the general review materials, which in the overt response versions greatly resembled quizzes--sequenced and prompted quizzes, with knowledge of correct response provided, but more difficult as programs than the initial instructional materials. Therefore the reading-only review and the massed practice possibly provided prompting and continuity which was useful to a certain extent for *everyone*. In examining this interaction in detail, for high ability students at least, the critical variable appears to be distribution of practice. It is possible that the more concerted effort allowed in the massed practice conditions was beneficial to the performance of high ability students--who might be expected to have longer attention span and therefore to suffer less from "learner fatigue" than would their less well-favored classmates.

In summarizing the results from both the overall gain from pre-acquisition to delayed retention and the percent correct on the delayed-retention test, the underlying assumptions concerning interactions of experimental treatments with ability were somewhat supported. The specific predictions of the order of these interactions were verified at least in part. It was predicted that high ability *S*s would learn proportionally more than would low ability *S*s from treatments hypothesized to be generally *less* effective. This prediction was realized to such an extent that in some instances high ability students actually learned more from these hypothetically poorer conditions (reading-only response mode and massed practice during review) than they learned from conditions predicted to produce generally superior learning (overt response mode and spaced review). The explanation for this could be postulated as caused by placement of materials along a sliding scale of suitability for students of different mental ability. That is, possibly these instruments (and the criterion tests accompanying them) were most effectively aimed for medium ability students. This would explain the better discrimination which was found among experimental conditions within medium ability level. Materials would, then, be presumed easier for higher ability students, and their own preference for the reading response mode and massed practice treatments which were more economical of time would have a positive influence on their learning from these.

For the specific prediction that low ability *S*s would profit proportionall more than would high ability *S*s from treatments hypothesized to be generally *more* effective the results of these analyses are not conclusive, since the amount learned was fairly small and since significant differences between treatments could not be demonstrated within the low ability level. However, from an examination of the rank order of treatments for low ability students, there appears to be some basis upon which to plan additional research which would better delineate such predicted differences.

Effects of Review.

Short-Term Effects of Review.

The short-term effects of review were ascertained by subtracting each S's post-acquisition test score from his corresponding post-review score. Thus a positive difference shows gain and a negative difference shows loss. Analyses were made of difference scores on the Long General Test and the Chronology Test.

The Short General Test was not given at the post-acquisition testing. It had been administered at the pre-acquisition testing time, only three weeks earlier at the most, and it was omitted at the post-acquisition testing time in order to avoid as much as possible any interactions of the instructional program with the measuring instrument. The Concepts Test was not made up until after the post-acquisition testing time, and was not administered at all until the post-review testing time. Short-term effects from review were, therefore, not available for either the Short General Test or the Concepts Test.

The Long General Test. Depending upon the groups examined, there was some gain and very small loss for the 15-week period between the post-acquisition and the post-review testing times, during which all students had received approximately five and one-third hours of either spaced or massed review instruction. The Long General Test indicated no discriminable main effects or interactions from experimental treatments, other than the significant main effect from Ss' level of ability ($p < .05$). (See analysis of variance, Table VII, Appendix B.) High Ability subjects had a mean gain of 5.04 percent, Middle Ability subjects had a mean gain of -1.80 percent, while Low Ability subjects had a mean gain of -.09 percent (an overall mean gain of 3.05%, see the graph in Figure 13).[2] These differences look very small, but for complex subject-matter such as this, the loss over a period of 15 weeks has often been found as high as 60% (Pressey and Robinson, 1944, pp. 540-549). The fact that the review procedure could maintain performance at so high a level indicates a success for the procedure, even though no discriminable differences could be found for specific experimental treatments.

The Chronology Test. The gain shown with the Chronology Test for the time from post-acquisition to post-review was higher than that shown with the Long General Test. The familiar significant effect owing to ability was present ($p < .005$), and mean gain for High Ability subjects was 12.7 percent, for Medium Ability subjects was 7.4 percent, and for Low Ability subjects was 5.8 percent. Table 7 shows the results of this analysis of variance (also in Table VIII, Appendix B). (See the graph, Figure 14.)

---

[2] These percent gain scores, as well as the percent scores shown in graphs are adjusted to accomodate differences between scores on multiple-choice and open-end test forms. This adjustment was necessary only for graphic purposes, not for analyses.

Table 7

Analysis of Variance indicating effects measured by the Chronology Test, from post-acquisition to post-review.

Source	df	MS	F	p
Text	1	.65		
Review	1	.96	1.31	
Practice	1	.91	1.24	
Ability	2	4.42	6.05	<.005*
T x R	1	.03		
T x P	1	.74	1.01	
T x A	2	.10		
R x P	1	.09		
R x A	2	1.75	2.39	<.10
P x A	2	4.29	5.87	<.005*
TxRxP	1	.43		
TxRxA	2	1.30	1.78	<.20
TxPxA	2	2.06	2.82	<.10
RxPxA	2	1.83	2.50	<.10
TxRxPxA	2	.44		
ERROR	90	.73		

*Indicates an effect accepted as reliable.

In addition to the main effect for ability, there was a significant interaction of distribution of practice during review, with level of ability (P x A; p <.005). Table 8 gives means and critical differences for the significant P x A interaction. Critical differences for p =.05 with this analysis is 1.06, and an (*) indicates a difference which reaches or exceeds the critical level. As the reader will see from his examination of Table 8, only a very cautious interpretation may be placed upon these effects. About the only safe conclusion which can be made is that Ss of medium ability seem to have profited from having spaced practice during review.

Table 8

Means and Critical Differences: Practice x Ability ($p < .005$), Chronology Test, Post-Acquisition to Post-Review.

Ability			High		Medium		Low	
	Practice		S	M	S	M	S	M
		Means	-.35	.02	2.24	.16	-.28	.27
High	S	-.35		-.37	-2.59*	-.51	-.07	-.62
	M	.02			-2.22*	-.14	.30	-.25
Medium	S	2.24				2.08*	2.52*	1.97*
	M	.16					.44	-.11
	S	-.28						-.55
	M	.27						

*$p < .05$

Long-Term Effects of Review.

The long-term effects of review were ascertained by subtracting each S's post-acquisition test score from his corresponding delayed retention test score. Thus a positive difference shows gain and a negative difference shows loss. Analyses were made of difference scores on the Long General Test and the Chronology Test. The time between post-acquisition and delayed-retention testing was thirty weeks, the last half of which was during summer vacation. Over this entire period of time, extremely small change scores were shown by the two measuring instruments. Neither of the analyses performed for this time period showed a significant main effect owing to ability. There were, however, a number of significant interactions.

The Long General Test. Table 9 shows the results of the analysis of variance for the Long General Test (also Table IX in Appendix B).

Table 9

Analysis of Variance indicating effects measured by the Long General Test, from post-acquisition to delayed-retention.

Source	df	MS	F	p
Text	1	3.72		
Review	1			
Practice	1	.36		
Ability	2	9.76	2.44	<.10
T x R	1	5.76	1.44	
T x P	1	.07		
T x A	2	1.28		
R x P	1			
R x A	2	1.76		
P x A	2	1.38		
TxRxP	1	16.74	4.18	<.05*
TxRxA	2	15.25	3.81	<.05*
TxPxA	2	1.70		
RxPxA	2	8.95	2.24	<.20
TxRxPxA	2	1.38		
ERROR	94	4.00		

* indicates a value accepted as reliable ($p < .05$)

Mean adjusted percent gain, as shown in Figure 14 for high ability students was -3.2, for medium ability students was -4., and for low ability students was -8., a mean loss of 5.4 percent over the thirty weeks between post-acquisition and delayed-retention testings, again a favorable result compared with retention studies of conventional materials.

Two significant interaction effects were found with the analysis of variance: response mode on text, by response mode on review, by distribution of practice (T x R x P; $p < .05$); and response mode on text, by response mode on review, by level of ability (T x R x A; $p < .05$). The critical difference for $p = .05$ for this analysis was 2.52. Means and critical differences for the T x R x P interaction are given in Table 10, and the rank-order expansion of the interaction is given in Figure 7. The type of practice (Spaced or Massed) was chosen to provide the organization of the experimental conditions, because this organization seemed to show the most interesting information.

Table 10
Means and Critical Differences, Text x Review x Practice Interaction ($p <.05$), Long General Test, Post-Acquisition to Delayed Retention.

Practice				Spaced Practice				Massed Practice			
	Text			OT		RT		OT		RT	
		Review		OR	RR	OR	RR	OR	RR	OR	RR
			Mean	-0.46	-1.03	-0.25	0.56	-1.72	0.93	1.61	-1.04
Spaced	OT	OR	-0.46		0.57	-0.21	-1.02	1.26	-1.39	-2.07	0.58
		RR	-1.03			-0.78	-1.59	0.69	-1.96	-2.64*	0.01
	RT	OR	-0.25				-0.81	1.47	-1.18	-1.86	0.79
		RR	0.56					2.28	-0.37	-1.05	1.60
Massed	OT	OR	-1.72						-2.65*	-3.33*	-0.68
		RR	0.93							-0.68	1.97
	RT	OR	1.61								2.65*
		RR	-1.04								

*$p <.05$

Figure 7
Rank Orders of Means: Text x Review x Practice Interaction ($p <.05$) Organized According to Practice, Long General Test, Post-Acquisition to Delayed-Retention

	Spaced Practice		Massed Practice	
	Overt Text	Reading Text	Overt Text	Reading Text
Overt Review	4.5 2.5	4.5 2.5	8 4	1 1
Reading Review	6 4	3 1	2 2	7 3

* Arrows indicate significant differences between treatments within each ability level. (⟵⟶) The numbers in the large boxes reflect the rank order of the treatments within each ability level. The number enclosed in the upper-right corner of each box reflects the position of that cell over the entire interaction. A rank of 1 indicates the best treatment.

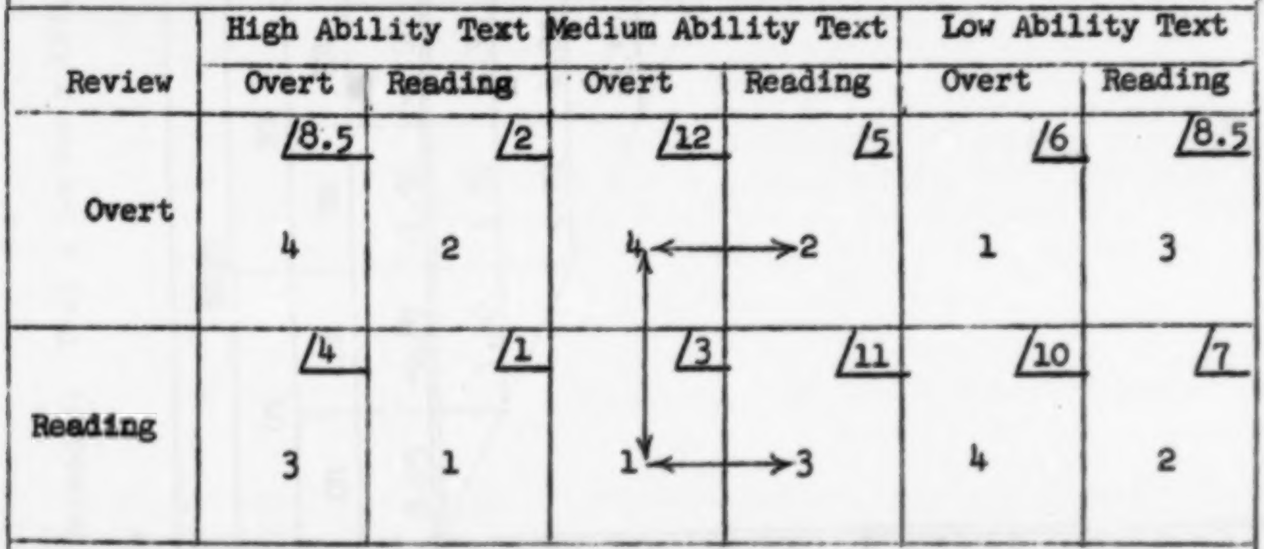
In this interaction, both of the conditions showing greatest mean gain and both of the conditions showing greatest mean loss involved Massed Practice. RT/OR/MP and OT/RR/MP were the two best conditions, ranked 1 and 2 overall; and OT/OR/MP and RT/RR/MP were the two poorest conditions, ranked 7 and 8 overall. This indicates that for students having massed practice on review, the crucial factor was the combination of response mode on text and response mode on review. Having a different response mode on review than on text was significantly superior to having the same response mode on both text and review. For students having spaced practice there were no additional determining factors. The spaced practice conditions are clustered together in the center of the interaction, having neither the very best nor the very worst conditions among them. The differences among the spaced practice conditions were not significant. Only the best ranked massed practice condition (RT/OR/MP, ranked 1 overall) was significantly discriminable from any spaced practice condition (OT/RR/SP, ranked 6 overall).

The present state of the "art" and "science" do not give us any method for or basis upon which these findings can be interpreted. If one were forced to recommend for other subject-matters, possibly the condition of least risk, Spaced Practice, would be preferable.

The Text x Review x Ability interaction was also found to be significant for this analysis (long-term effects of review, as shown by the Long General Test). Table 11 contains means and critical differences, and Figure 8 contains the rank ordering of these interaction means.

Figure 8

Rank Orders: Text x Review x Ability Interaction Treatments of Long General Test, Post-Acquisition less Delayed-Retention ($p < .05$)



	High Ability Text		Medium Ability Text		Low Ability Text	
Review	Overt	Reading	Overt	Reading	Overt	Reading
Overt	8.5 4	2 2	12 4	5 2	6 1	8.5 3
Reading	4 3	1 1	3 1	11 3	10 4	7 2

Arrows (←→) indicate significant within-level treatment differences. The numbers in the large boxes reflect the rank order of the treatments within each ability level. The number enclosed in the upper-right corner of each box reflects the position of that cell over the entire interaction. A rank of 1 indicates the best treatment.

Table 11

Means and Critical Differences: Text x Review x Ability Interaction (p <.05), Long General Test, Post-Acquisition less Delayed-Retention.

Ability				High				Medium				Low			
	Text			OT		RT		OT		RT		OT		RT	
		Review		OR	RR	OR	RR	OR	RR	OR	RR	OR	RR	OR	RR
			Mean	0.25	-0.85	-1.56	-2.08	3.23	-1.15	-0.72	2.92	-0.22	2.16	0.25	-0.14
High	OT	OR	0.25		1.10	1.81	2.33	-2.98*	1.40	0.97	-2.67*	0.47	-1.91	0.00	0.39
		RR	-0.85			0.71	1.23	-4.08*	0.30	-0.13	-3.77*	-0.63	-3.01*	-1.10	-0.71
	RT	OR	-1.56				0.52	-4.79*	-0.41	-0.84	-4.48*	-1.34	-3.72*	-1.81	-1.42
		RR	-2.08					-5.31*	-0.93	-1.36	-5.00*	-1.86	-4.24*	-2.33	-1.94
Medium	OT	OR	3.23						4.38*	3.95*	0.31	3.45*	1.07	2.98*	3.37*
		RR	-1.15							-0.43	-4.07*	-0.93	-3.31*	-1.40	-1.01
	RT	OR	-0.72								-2.20*	-0.50	-2.88*	-0.97	-0.58
		RR	2.92									3.14*	0.76	2.67	3.06*
Low	OT	OR	-0.22										-2.38	-0.47	-0.08
		RR	2.16											1.91	2.30
	RT	OR	0.25												0.39
		RR	-0.14												

*p <.05 (A negative mean shows gain from post-acquisition to delayed-retention.)

In the T x R x A interaction there were significant differences *between* all ability levels; however, treatment differences within ability levels appeared only for medium ability Ss. The rank orders show that for medium ability students the treatment MA/OT/RR (ranked 3 overall) produced the highest mean gain, and the treatment MA/OT/OR (ranked 12 overall) produced the greatest mean loss. The crucial factor here appeared to be response mode on review. For medium ability students who had overt response versions of the initial instructional text, the reading-only review produced significantly higher gain than did overt response review.

Overall, the reading response mode on the initial text (RT) produced somewhat higher gain scores, although not without exception, and the reading response mode on review (RR) produced somewhat higher gain scores, although again not without exception. The rankings of conditions *within* ability levels gave some support to the hypothesis that high ability students would have proportionally higher scores than would lower ability students, from treatments hypothesized to be generally less effective--those involving reading-only response mode on the initial text (RT) and reading-only response mode on review (RR). Again, the evidence is very ambiguous for the similar hypothesis that lower ability students would have proportionally greater gain scores than would higher ability students, from treatments hypothesized to be generally more effective (OT and OR).

The Chronology Test. The Chronology Test was compared with the Long General Test for the same gain scores (post-acquisition to delayed-retention), and similarities and differences are noted below. There are several instances of similarity between effects shown by the two different measuring instruments. The similarities in non-reliable findings are not very interesting outside of the fact that neither the Long General Test nor the Chronology Test showed a reliable effect owing to ability. Mean adjusted percent gain on the Chronolog Test was 7.90 for high ability, 1.15 for medium ability, and .88 for low ability students, a mean percent gain of 3.31.

The most interesting similarity beteen effects shown by the two measuring instruments was in the significant T x R x A interaction which appeared in both analyses. The most interesting difference is in the main effect (as predicted) for response mode on review, which appears in the analysis of the Chronology Test.

Table 12 (also Table X in Appendix B) gives the results from the analysis of variance performed with these gain scores from the Chronology Test. Critical difference for $p = .05$ is 1.07.

Following down the analysis of variance table, we first come to the significant main effect ($p < .01$) in favor of overt response mode on review (mean differences: OR = .66; RR = -.33). Next we find the significant interaction ($p < .05$) of response mode on review, with schedule of practice on review (R x P). Table 13 gives means and critical differences for this interaction.

Table 12

Analysis of Variance indicating effects measured by the Chronology test, from post-acquisition to delayed-retention.

Source	df	MS	F	p
Text	1	.01		
Review	1	5.88	7.95	<.01*
Practice	1	2.23	3.01	<.10
Ability	2	1.14	1.50	
T x R	1	.02		
T x P	1	1.08	1.46	
T x A	2	.80	1.08	
R x P	1	3.11	4.20	<.05*
R x A	2	.33		
P x A	2	2.04	2.76	<.10
T x R x P	1	.04		
T x R x A	2	3.74	5.05	<.01*
T x P x A	2	.56		
R x P x A	2	2.80	3.78	<.05*
TxRxPxA	2	.47		
ERROR	92	.74		

*indicates effect accepted as reliable (p < .05).

Table 13

Means and Critical Differences: Review x Practice Interaction (p <.05), Chronology Test, Post-Acquisition to Delayed-Retention.

Review			OR		RR	
	Practice		SP	MP	SP	MP.
		mean	.61	.72	.33	-1.00
OR	SP	.61		-.11	.28	1.61*
	MP	.72			.39	1.72*
RR	SP	.33				1.33*
	MP	-1.00				

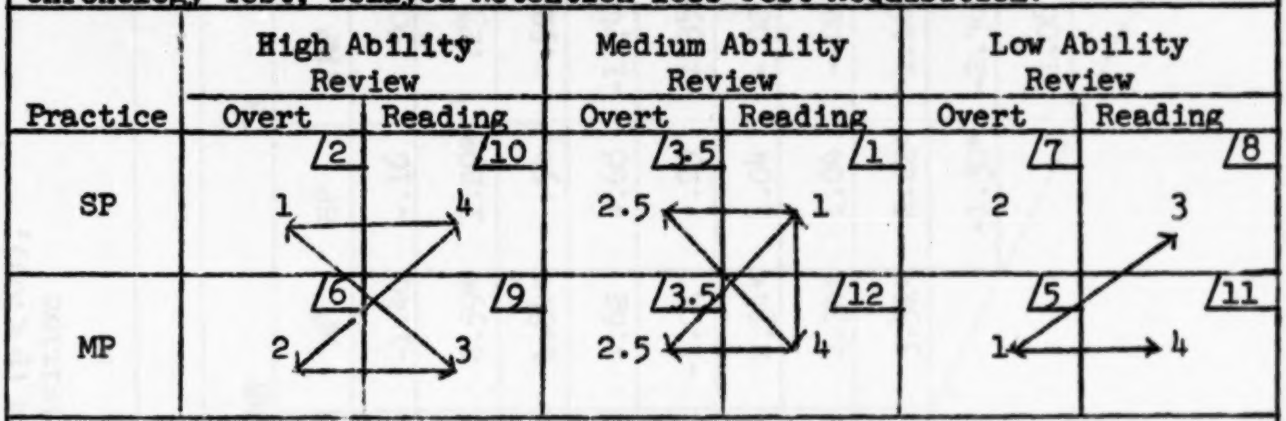
* p <.05

We see from Table 13 that the significant effect seems to be produced by only one condition (RR/MP) being significantly poorer than all others, but none of the other conditions can be discriminated from one another. Possibly, in this instance, the conclusion may be drawn that with reading-only review, a massed practice condition would be detrimental to retention. It is not, however, until the two significant three-way interactions (T x R x A, p< .01; and P x R x A, p<.05) are examined that any clear interpretation can be placed on the significant main effect and two-way interation.

The means and critical differences for the practice with review with ability interaction (P x R x A) are shown in Table 14, and the means and critical differences for the text with review with ability interaction, (T x R x A) are shown in Table 15. The rank orders for these two three-way interactions are given in Figures 9 and 10.

Figure 9

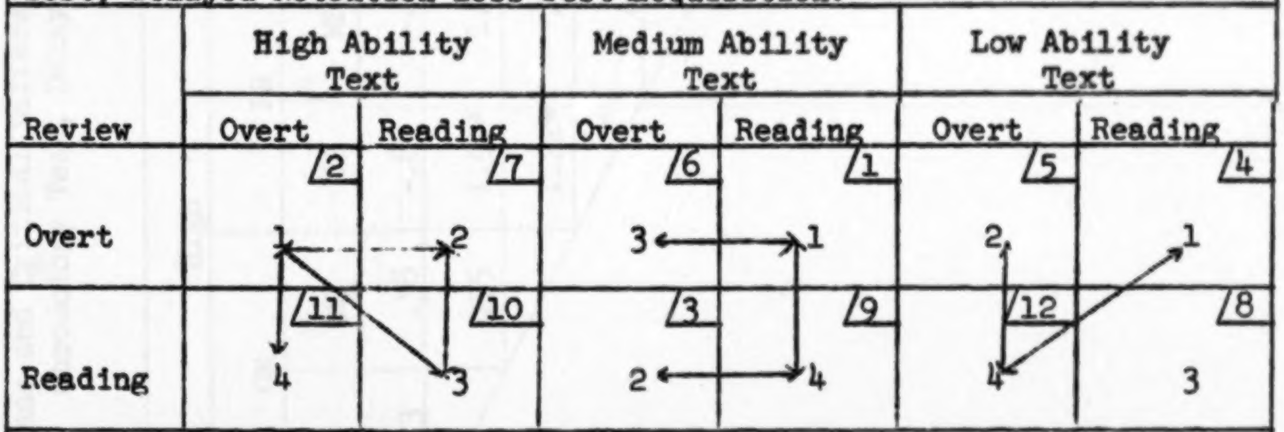
Rank Orders: Practice x Review x Ability Interaction ($p < .05$), Chronology Test, Delayed-Retention less Post-Acquisition.


High Ability
Review
Medium Ability
Review
Low Ability
Review
Practice
Overt
Reading
Overt
Reading
Overt
Reading
2
10
3.5
1
7
8
SP
1
4
2.5
1
2
3
6
9
3.5
12
5
11
MP
2
3
2.5
4
1
4

Arrows indicate significant within-level treatment differences. The numbers in the large boxes reflect the rank order of the treatments within each ability level. The number enclosed in the upper right corner of each box reflects the position of that cell over the entire interaction. A rank of 1 indicates the best treatment.

Figure 10

Rank Orders: Text x Review x Ability Interaction ($p < .01$), Chronology Test, Delayed-Retention less Post-Acquisition.


High Ability
Text
Medium Ability
Text
Low Ability
Text
Review
Overt
Reading
Overt
Reading
Overt
Reading
2
7
6
1
5
4
Overt
1
2
3
1
2
1
11
10
3
9
12
8
Reading
4
3
2
4
4
3

Arrows indicate significant differences between treatments within each ability level. (⟷) The numbers in the large boxes reflect the rank order of the treatments within each ability level. The number enclosed in the upper right corner of each box reflects the position of that cell over the entire interaction. A rank of 1 indicates the best treatment.

Table 14

Means and Critical Differences: P x R x A Interaction ($p < .05$), Chronology Test, Delayed Retention less Post-Acquisition

Ability				High				Medium				Low			
	Review			OR		RR		OR		RR		OR		RR	
		Practice		SP	MP	SP	MP	SP	MP	SP	MP	SP	MP	SP	MP
			Means	1.13	.38	-.84	-.45	.88	.88	2.06	-1.46	-.16	.90	-.23	-1.08
High	OR	SP	1.13		.75	1.97*	1.58*	.25	.25	-.93	2.59*	1.29*	.23	1.36*	2.21*
		MP	.38			1.22*	.83	-.50	-.50	1.68*	1.84*	.54	-.52	.61	1.46*
	RR	SP	-.84				.39	-1.72*	-1.72*	-2.90*	-.62	-.68	-1.74*	-.61	-.24
		MP	-.45					-1.33*	-1.33*	-2.51*	-1.01	-.29	-1.35*	-.22	-.63
Medium	OR	SP	.88						.00	1.18*	2.34*	1.04	-.02	1.11*	1.96*
		MP	.88							1.18*	2.34*	1.04	-.02	1.11*	1.96*
	RR	SP	2.06								3.52*	2.22*	1.16*	2.29*	3.14*
		MP	-1.46									-1.30*	-2.36*	-1.23*	-.38
Low	OR	SP	-.16										-1.06	.07	.92
		MP	.90											1.13*	1.98*
	RR	SP	-.23												-.85
		MP	-1.08												

*$p < .05$

Table 15

Means and Critical Differences: T x R x A Interaction ($p < .01$), Chronology Test, Delayed Retention less Post-Acquisition

Ability				High				Medium				Low			
	Review			OR		RR		OR		RR		OR		RR	
		Text		OT	RT	OT	RT	OT	RT	OT	RT	OT	RT	OT	RT
			Mean	1.56	-.06	-.78	-.50	-.04	1.79	1.00	-.40	.34	.40	-1.18	-.14
High	OR	OT	1.56		1.62	2.34	2.06	1.60	-.23	.56	1.96	1.22	1.16	2.74*	1.90
		RT	-.06			.72	.44	-.02	1.85*	-1.06	.34	-.40	.46	1.12*	.08
	RR	OT	-.78				-.28	-.74	2.57*	1.78*	-.38	1.12*	1.18	-.40	.64
		RT	-.50					-.46	2.29*	1.50*	-.10	-.84	-.90	.68	-.36
Medium	OR	OT	-.04						-1.75*	1.04	-.36	-.38	-.44	1.14*	.10
		RT	1.79							.79	2.19*	1.45*	1.39*	2.97*	1.93*
	RR	OT	1.00								1.40*	.66	.60	2.18*	1.14*
		RT	-.40									.74	-.80	-.78	-.26
Low	OR	OT	.34										-.06	1.52*	.48
		RT	.40											1.58*	.54
	RR	OT	-1.18												1.04
		RT	-.14												

*$p < .05$

With careful examination of these two three-way interactions, it is somewhat easier to interpret the entire analysis. In general, Overt Review was "better" than Reading Review--with an odd exception occurring in the medium ability level: in the P x R x A interaction the treatment of rank order 1 within medium ability was a Reading Review condition (SP/RR). This condition was also rank order 1 for the entire interaction, and differed significantly from each other condition except for HA/OR/SP. The condition MA/RR/MP, ranked 4 for medium ability students, was rank 12 for the entire interaction. This condition was poorer than all others excepting only HA/RR/MP, HA/RR/SP (ranked 9 and 10 overall) and LA/RR/MP (ranked 11 overall) from which MA/RR/MP did not discriminate. If one were to make recommendations as to the best method for learning these paired-associates-type chronological facts, a condition involving Overt Review and Spaced Practice would do the most good and the least harm for a heterogeneous population of the sort involved in this experiment.

In the T x R x A interaction the medium ability rank 1 was again rank 1 for the entire interaction, but this time the effect favored Overt Review (RT/OR). This condition was "better" than any other treatment cell excepting HA/OT/OR (ranked 2 overall) and MA/OT/RR (ranked 3 overall), from which conditions MA/RT/OR did not differ significantly. The cell ranked 4 for low ability Ss (OT/RR) was ranked 12 for the entire interaction. This condition was ranked significantly lower than each other cell in the interaction, excepting LA/RT/RR (ranked 8 overall), MA/RT/RR (ranked 9 overall), and HA/RT/RR and HA/OT/RR (ranked 10 and 11 overall), from which LA/OT/RR was not significantly different. It should be pointed out that for high ability students the reading review conditions were so poor as to make them two of the three lowest in the entire interaction, and that three of the treatment conditions for low ability students are of higher rank than the two poorest conditions for high ability students.

Again, if one were to make recommendations for Chronological and other paired-associates types of information, a condition involving overt response on review would seem clearly indicated. It is doubtful that any reasonable recommendation may be made at this point about the response mode on the initial instructional text. The reason for this is that the direct effects of the response mode employed during initial instruction were not taken into account specifically in this analysis, since the gain scores were obtained with post-acquisition as the starting point.

Summary of the Comparison. To summarize the comparison of the long-term effects of review (post-acquisition to delayed-retention difference scores; for the Long General and Chronology Tests): 1. No main effects were found owing to ability with either test. 2. No main effects at all were found with the Long General Test. 3. A significant main effect, in favor of having had overt response mode for chronological review materials, was found with the Chronology Test. 4. Several significant interactions were found with each test. 5. A significant T x R x A interaction was found for both tests. The T x R x A interaction indicated generally for the Chronology Test the favorability of overt response mode on chronological review for all ability levels. But the T x R x A interaction for the Long General Test, indicated some favorability for reading-only response for high and medium ability students on review of general materials.

Retention Losses During Periods of No Treatment.

The analyses of change where no treatment intervened between the post-review and delayed-retention measurements were made with the Long General Test, the Chronology Test, and the Concepts Test. The analysis of variance tables showing the results of these analyses are in Appendix B (Tables XI, XII, and XIII). These analyses verified the hypothesis that no main effect owing to ability would be found with difference scores, when only time (and no experimental treatment) intervened between any two testing points. In the present experiment, a summer vacation intervened between the post-review and the delayed-retention testings.

No main effects and no interaction effects whatsoever were observed with the Chronology and the Concepts Tests. Two rather interesting interactions were, however, found with the Long General Test: response mode on initial instructional text, by response mode on review (T x R; $p < .05$); and response mode on text, by response mode on review, by level of ability (T x R x A; $p < .025$). Results of this analysis of variance are in Table 16 (Table XI in Appendix B). Critical difference for this analysis is 2.04. Since differences between conditions in the Text by Review interaction were very small, reference must be made to the three-way interaction for an interpretation of results. Table 17 and Figure 11 below give means and critical differences, and rank ordering of the means for this interaction.

Table 16

Analysis of Variance indicating effects measured by the Long General Test, for the passage of time from Post-Review to Delayed-Retention.

Source	df	MS	F	p
Text	1	.02		
Review	1	.38		
Practice	1	1.22		
Ability	2	4.18	1.61	
T x R	1	13.41	5.18	<.05*
T x P	1	.00		
T x A	2	.52		
R x P	1	.00		
R x A	2	1.20		
P x A	2	1.20		
TxRxP	1	1.90		
TxRxA	2	10.58	4.08	<.025*
TxPxA	2	.30		
RxPxA	2	5.84	2.25	<.20
TxRxPxA	2	3.02	1.17	
ERROR	91	2.59		

*Indicates effect accepted as reliable ($p < .05$).

Figure 11

Rank Orders: Text x Review x Ability Interaction, Long General Test, Post-Review less Delayed-Retention ($p < .025$)

Review	High Ability Text		Medium Ability Text		Low Ability Text	
	Overt	Reading	Overt	Reading	Overt	Reading
Overt	$\boxed{6}$ 3	$\boxed{10}$ 4	$\boxed{9}$ 3 ↔ ↕	$\boxed{1}$ ↔ 1 ↕	$\boxed{11.5}$ 4	$\boxed{7}$ 2
Reading	$\boxed{4}$ 1	$\boxed{5}$ 2	$\boxed{2}$ ↕ 2 ↔	$\boxed{11.5}$ ↕ ↔ 4	$\boxed{3}$ 1	$\boxed{8}$ 3

An arrow (↔) indicates significant within-level treatment differences. The numbers in the large boxes reflect the rank order of the treatments within each ability level. The number enclosed in the upper-right corner of each box reflects the position of that cell over the entire interaction. A rank of 1 indicates the lowest mean loss.

Contrary to what might have been expected some treatment groups (MA/RT/OR and MA/OT/RR) actually showed an improvement between the post-review and the delayed-retention testing (the positive values in Table 17). The reason for this improvement is not clear and might be ascribed either to residual effects from the experimental treatments or possibly to uncontrolled variables, such as learning opportunity over the summer months and/or individual motivational differences. E. B. Parker has suggested that the instruction may have stimulated students to maintain an interest in the subject-matter and to pay attention to relevant information in their general environment.

It is very clear from the rank ordering of conditions shown in Figure 11 why no main effect was discriminable owing to ability. For example, the condition ranked 1 was within the medium ability level, the condition ranked 3 was within the low ability level, and the condition ranked 10 was within the high ability level.

The medium ability group with the RT/OR treatment was ranked 1 overall. That treatment was, however, the poorest for high ability students (ranked 10 overall). For high ability students RT/RR was not significantly poorer than any other, but for middle ability students RT/RR produced the greatest loss of all and was, in fact, one of the poorest conditions in the entire interacti

Table 17

Means and Critical Differences: Text x Review x Ability Interaction ($p < .025$), Long General Test, Post-Review to Delayed-Retention.

Ability				High				Medium				Low			
	Text			OT		RT		OT		RT		OT		RT	
		Review		OR	RR	OR	RR	OR	RR	OR	RR	OR	RR	OR	RR
			Mean	-1.44	-1.10	-2.52	-1.26	-2.32	1.08	2.06	-2.60	-2.60	-1.09	-1.55	-1.87
High	OT	OR	-1.44		-0.34	1.08	-0.18	0.88	-2.52*	-3.50*	1.16	1.16	-0.35	0.11	0.43
		RR	-1.10			1.42	0.16	1.22	-2.18*	-3.16*	1.50	1.50	-0.01	0.45	0.77
	RT	OR	-2.52				-1.26	-0.20	-3.60*	-4.58*	0.08	0.08	-1.43	-0.97	-0.65
		RR	-1.26					1.06	-2.34*	-3.32*	1.34	1.34	-0.17	0.29	0.61
Medium	OT	OR	-2.32						-3.40*	-4.38*	0.28	0.28	-1.23	-0.77	-0.45
		RR	1.08							-0.98	3.68*	3.68*	2.17*	2.63*	2.95*
	RT	OR	2.06								4.66*	4.66*	3.15*	3.61*	3.93*
		RR	-2.60									0.00	-1.51	-1.05	-0.73
Low	OT	OR	-2.60										-1.51	-1.05	-0.73
		RR	-1.09											0.46	0.78
	RT	OR	-1.55												0.32
		RR	-1.87												

*$p < .05$

In this T x R x A interaction, treatment group differences within Ability level were again focused on medium ability. However, while none of the differences were
many differences were significant between the treatment groups for different ability levels.

These results point up an interesting tendency found in several analyses for the middle ability students. This tendency was for the middle ability level to contain at least one treatment which showed greater gain than any--even any in the high ability level--and possibly to contain also one treatment which showed as small gain as any in the low ability level.

Further Examination of Results.

Ability and Differential Gain.

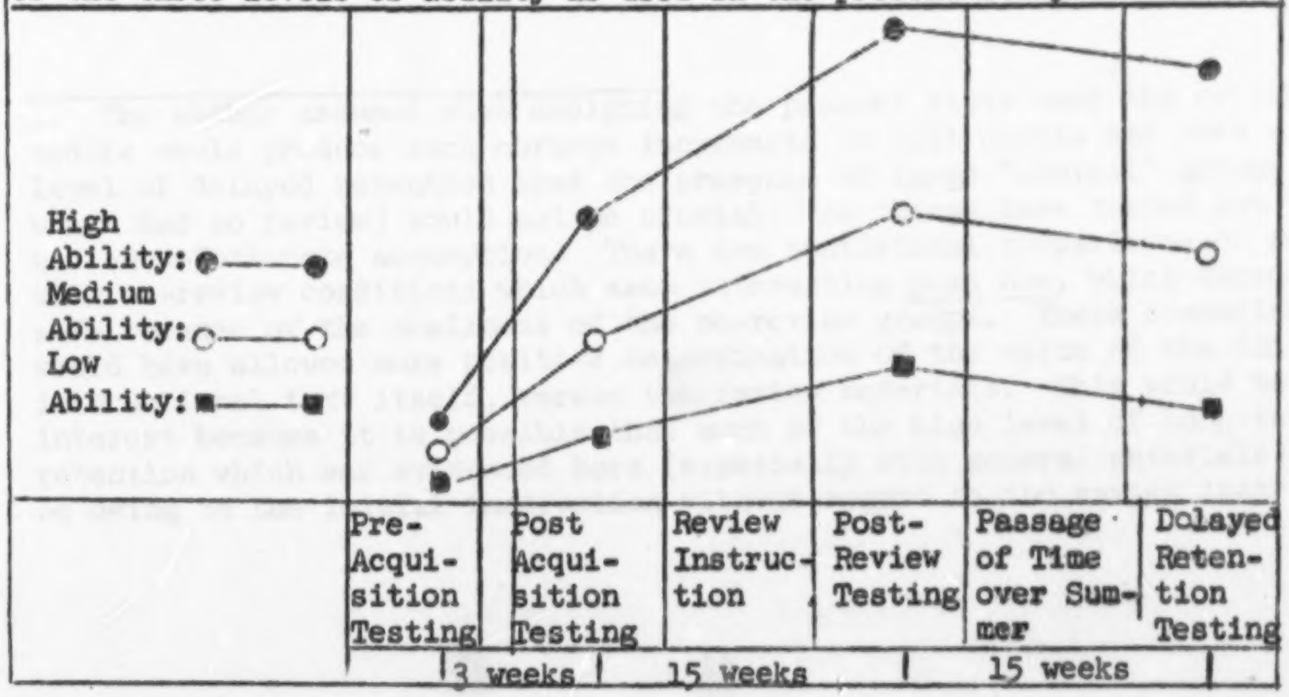
A significant main effect owing to ability was expected from all analyses of obtained correct scores. The analyses have confirmed this hypothesis.

A differential effect owing to ability was expected to be reflected in the slopes of lines connecting test points (i.e., in the analyses of difference scores) which were derived where some experimental treatment intervened between the beginning and ending points of the difference. This differential effect was expected to produce lines of greater slope for students of higher ability.

No significant effect from ability was expected on slopes of lines connecting testing points (difference scores) over the summer months, between post-review and delayed-retention where time was the only "treatment" intervening.

Figure 12 below shows the hypothetical main effects and differential effects expected owing to ability.

Figure 12
A theoretical diagram showing hypothesized results of analyses, in terms of the three levels of ability as used in the present study.



In terms of only the three levels of ability, disregarding specific treatments, this hypothesis was substantiated fairly consistently by the analyses. On the following pages are diagrams showing actual effects of instruction, according to ability levels. Included in these diagrams are the "curves" for the three small groups of Ss (high, medium, and low ability) who, for various reasons, were not assigned to any review condition.[1] These Ss were a random selection of the population of each of the three ability levels. However, because of the very small numbers of Ss the data from these no-review conditions could not be sensibly analyzed in comparison to the larger experimental group. Also, due to the effect on the variance, of a fairly wide score range and a small N, these no-review conditions were not satisfactorily compared among themselves ($F<1.00$, in each case). Their data would, in fact, not be included here, if it were not extremely interesting. The reliability of the sampling is possibly demonstrated by the close correspondence between the results reflected by these small no-review groups and the expectations one might have of how these groups should compare to the larger experimental population.

A statistical comparison of these various line slopes would have involved making a great many questionable assumptions. For this reason the following graphic representations will be discussed and compared without offering statistical evidence of differences within them.

The Long General Test. The first diagram (Figure 13) presents connected test points, obtained from the Long General Test, with points shown at the pre-acquisition time being supplied from the Short General Test (which the reader will recall was sampled from the same general item pool as was the Long General Test). In this diagram, neither the Middle Ability level nor the Low Ability level can be considered to have satisfied the hypothetical model (lines should have positive slope, showing gain due to review), although the losses were very small from both of these groups over that period of time.

---

[1] The author assumed when designing the present study that the review procedure would produce such obvious increments in test scores and such a high level of delayed retention that the presence of large "control" groups (Ss which had no review) would not be crucial. As things have turned out this was an unfortunate assumption. There are statistical comparisons of review with no-review conditions which seem interesting post hoc, which cannot be made because of the smallness of the no-review groups. These comparisons would have allowed more positive determination of the value of the initial instructional text itself, versus the review materials. This would be of interest because it is possible that much of the high level of long-term retention which was evidenced here (especially with general materials) could be owing to the initial instruction without regard to the review instruction.

Figure 13

Mean per cent correct on the Long General Test for three levels of ability.

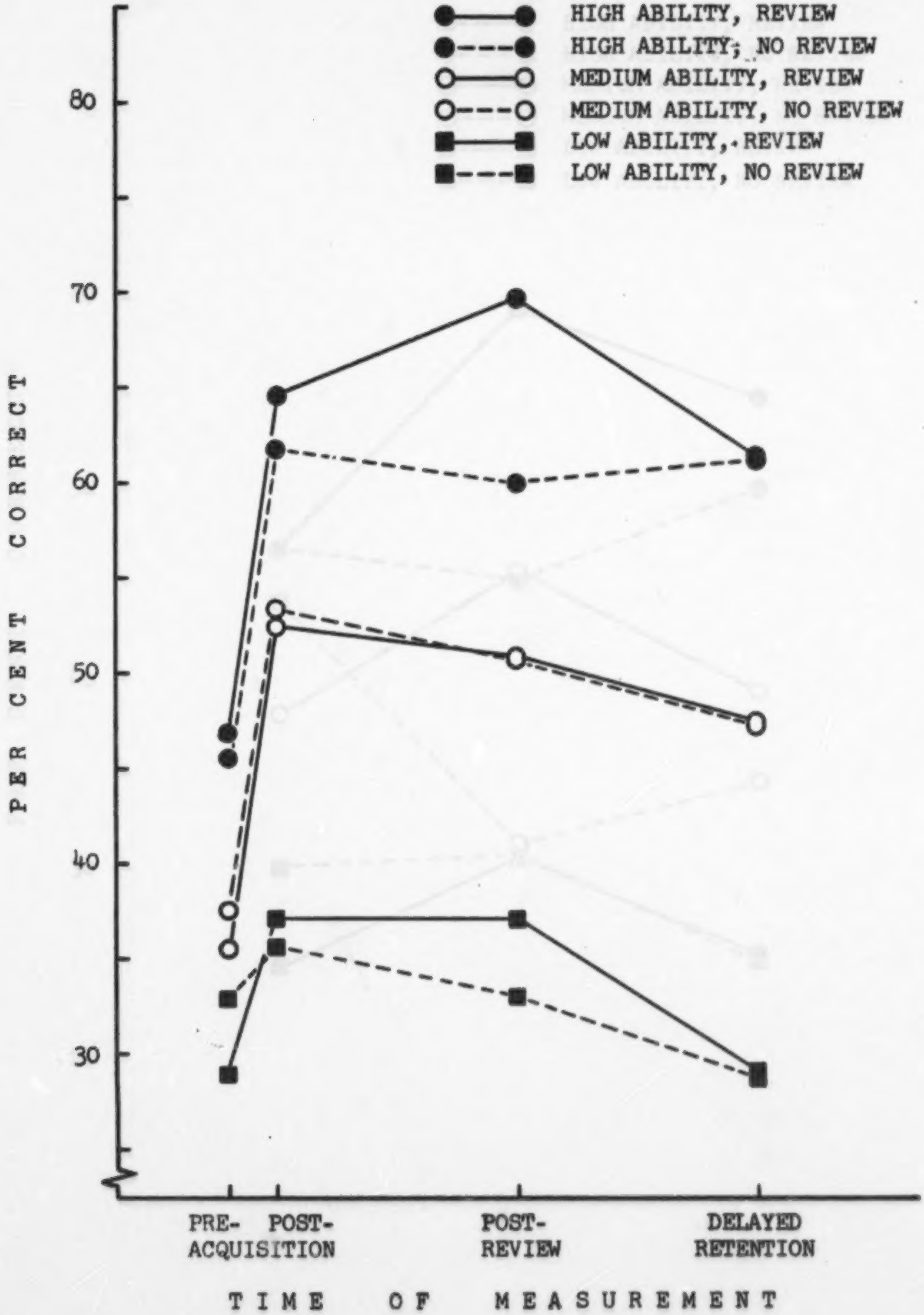


Figure 14
Mean per cent correct on the Chronology Test for three levels of ability.

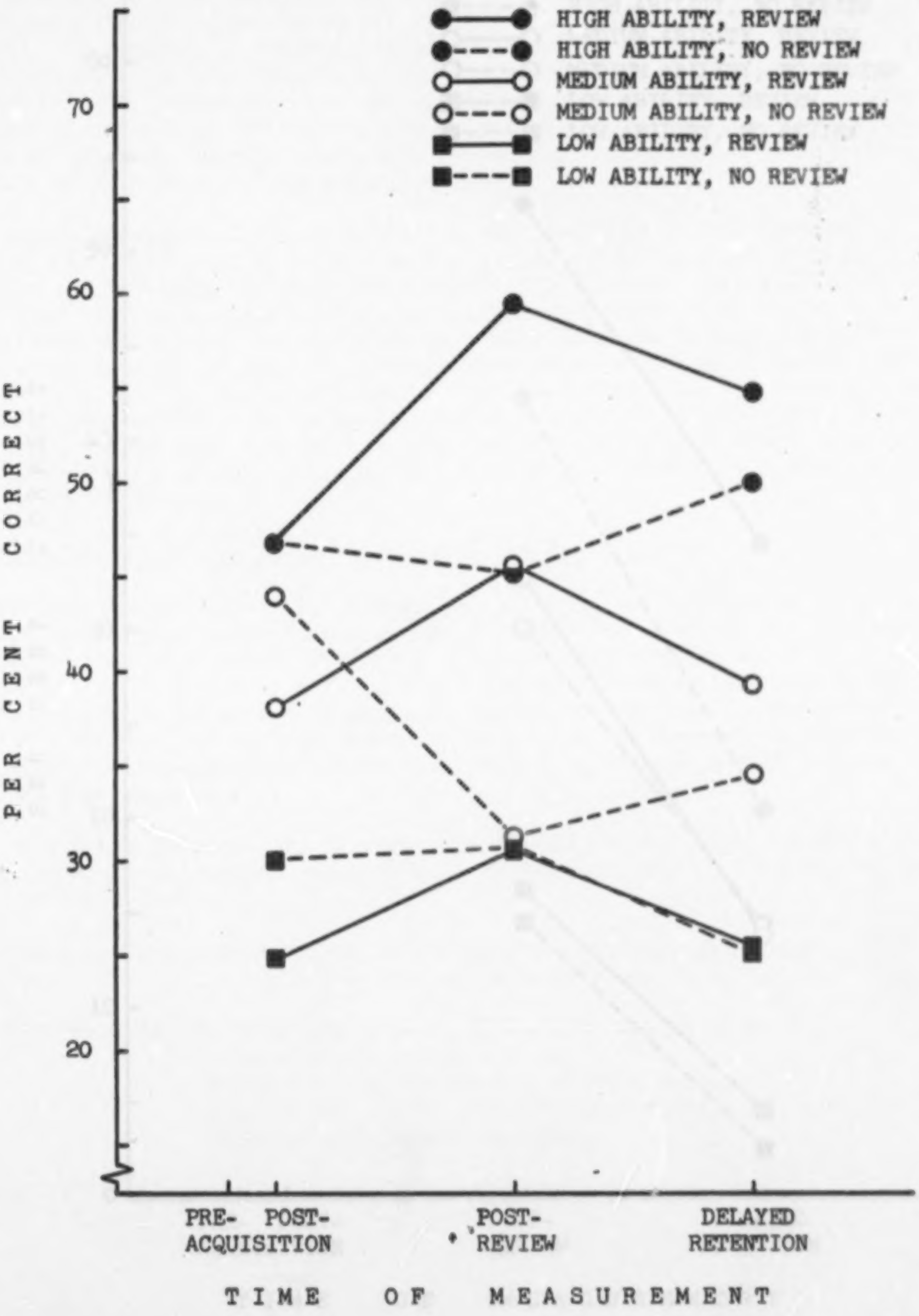
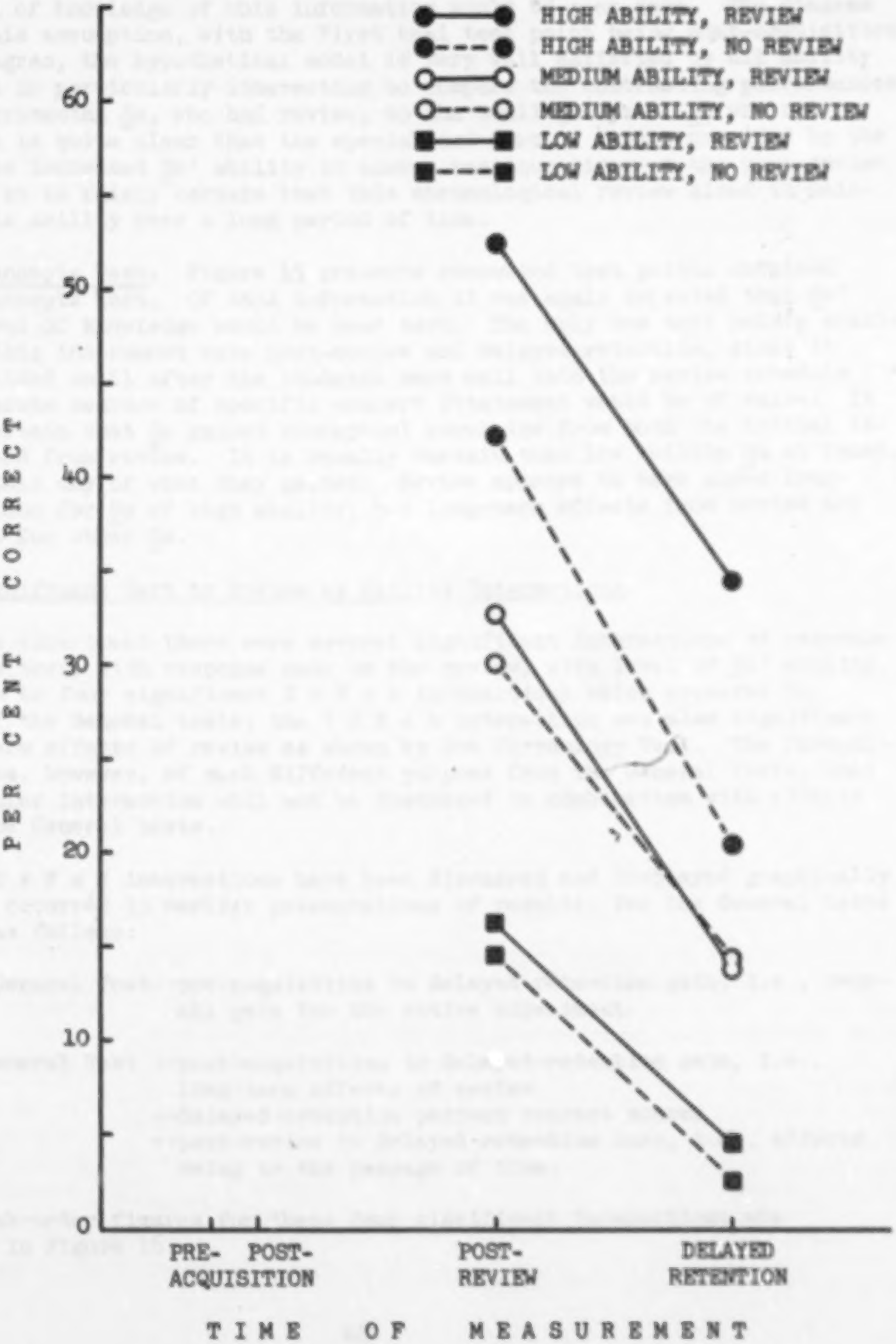


Figure 15

Mean per cent correct on the Concepts Test for three levels of ability.



The Chronology Test. The next diagram (Figure 14) presents connected test points, obtained from the Chronology Test. It was expected that Ss' initial level of knowledge of this information would be near zero. The diagram reflects this assumption, with the first real test point being post-acquisition. In this diagram, the hypothetical model is very well satisfied by all ability levels. It is particularly interesting to compare the contrasting performances of the experimental Ss, who had review, to the small groups of Ss who had no review. It is quite clear that the specialized form of review provided by the chronologies increased Ss' ability to answer test questions at the post-review level, and it is fairly certain that this chronological review aided in maintaining this ability over a long period of time.

The Concepts Test. Figure 15 presents connected test points obtained from the Concepts Test. Of this information it was again expected that Ss' initial level of knowledge would be near zero. The only two test points available with this instrument were post-review and delayed-retention, since it was not decided until after the students were well into the review schedule that a separate measure of specific concept attainment would be of value. It is fairly certain that Ss gained conceptual knowledge from both the initial instruction and from review. It is equally certain that Low Ability Ss, at least, did not retain any of what they gained. Review appears to have aided long-term retention for Ss of high ability, but long-term effects from review are not obvious for other Ss.

The Significant Text by Review by Ability Interactions.

In this experiment there were several significant interactions of response mode on the text, with response mode on the review, with level of Ss' ability. In addition to four significant T x R x A interactions which appeared in analyses of the General tests, the T x R x A interaction was also significant for long-term effects of review as shown by the Chronology Test. The Chronology Test was, however, of such different purpose from the General tests, that the particular interaction will not be discussed in combination with effects shown on the General tests.

These T x R x A interactions have been discussed and displayed graphically where they occurred in earlier presentations of results; for the General tests they were as follows:

Short General Test--pre-acquisition to delayed-retention gain, i.e., overall gain for the entire experiment.

Long General Test --post-acquisition to delayed-retention gain, i.e., long-term effects of review
--delayed-retention percent correct scores
--post-review to delayed-retention loss, i.e., effects owing to the passage of time.

The rank-order figures for these four significant interactions are reproduced in Figure 16.

Figure 16

Four significant interactions of response mode on initial text, with response mode on review, with students' level of ability (T x R x A).

Identification of Interaction	Response Mode on Review	High Ability		Medium Ability		Low Ability	
		Overt Text	Reading Text	Overt Text	Reading Text	Overt Text	Reading Text
Short General Test, pre-acquisition to delayed-retention	Overt	2 ↑ /4	3 /5	4 ↑ ← /8	→ 2 /3	1 /7	3 /11
	Reading	4 ↓ ← /9	→ 1 /2	1 ↓ ← /1	→ 3 /6	2 /10	4 /12
Long General Test, post-acquisition to delayed-retention	Overt	4 /8.5	2 /2	4 ↑ ← /12	→ 2 /5	1 /6	3 /8.5
	Reading	3 /4	1 /1	1 ↓ ← /3	→ 3 /11	4 /10	2 /7
Long General Test, delayed-retention raw scores	Overt	2 /2	4 ↑ /5	3 ↑ /7	↗ 2 /6	2 /10	3 /11
	Reading	3 /3	1 ↓ /1	1 ↓ ↙ ← /4	→ 4 /8	1 /9	4 /12
Long General Test, post-review to delayed-retention	Overt	3 /6	4 /10	3 ↑ ← /9	→ 1 ↑ /1	4 /11.5	2 /7
	Reading	1 /4	2 /5	2 ↓ ← /2	→ 4 ↓ /11.5	1 /3	3 /8

Arrows indicate significant differences between treatments within each ability level (←→). The numbers in the large boxes reflect the rank order of the treatments within each ability level. The number enclosed in the upper-right corner of each box reflects the position of that cell over the entire interaction. A rank of 1 indicates the best treatment.

Composite Average* of all the above Interactions	Response Mode on Review	High Ability Overt Text	High Ability Reading Text	Medium Ability Overt Text	Medium Ability Reading Text	Low Ability Overt Text	Low Ability Reading Text
	Overt	2.75 /5.12	3.25 /5.5	3.50 /9.00	1.75 /3.75	2.00 /8.60	2.75 /9.38
	Reading	2.75 /5.00	1.25 /2.25	1.25 /2.50	3.50 /9.12	2.00 /8.00	3.25 /9.75

*No significance tests were made with these composite average data.

At the very bottom of Figure 16 is a composite average of all the four interactions, and Figure 17 gives the mean composite rank for each part of each variable (summing over the other variable).

Figure 17

Composite mean ranks over each variable for each ability level, from the four significant T x R x A interactions shown in Figure 16.

High Ability		Medium Ability		Low Ability	
OR	5.31 3.00	OR	6.38 2.62	OR	8.99 2.38
RR	3.62 2.00	RR	5.81 2.38	RR	8.88 2.62
OT	RT	OT	RT	OT	RT
5.06 2.75	3.88 2.25	5.75 2.28	6.44 2.62	8.30 2.00	9.56 3.00

These composites show consistencies of effects with General tests over general materials over each of the analyses made for the duration of the experiment. These composites indicate that reading-only response mode was superior for high ability students on both initial text and on review. The composites also indicate that medium ability students profited more from a different response mode on review than they had on initial text, and some slight superiority is indicated for the mixture of overt response on initial text and reading-only response on review. For low ability students the composites indicate that overt responding on both initial text and review materials was superior.

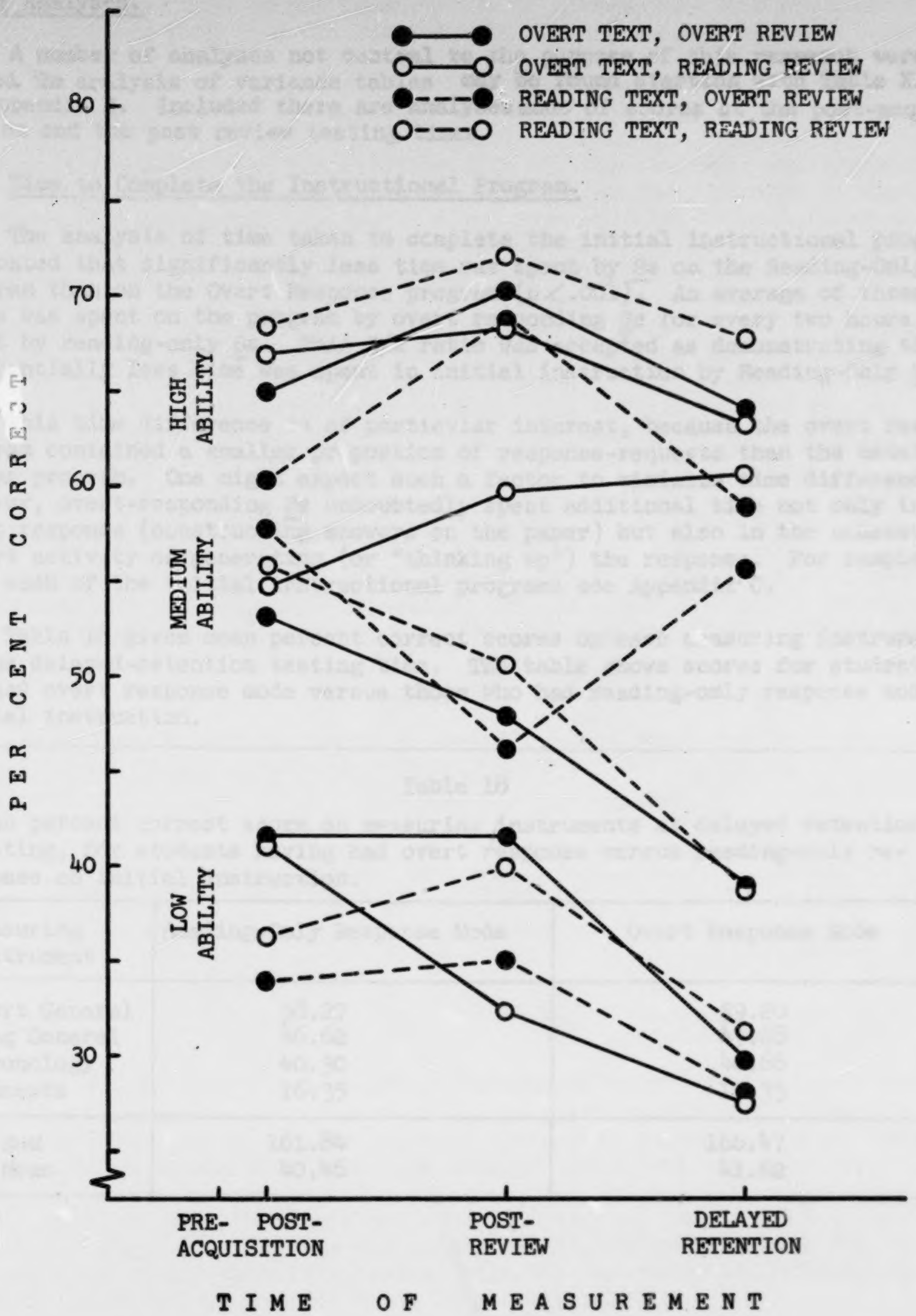
The hypotheses concerning interactions of experimental conditions with students' level of ability were (a) higher ability students would be able to learn proportionally more than would lower ability students, from experimental treatments predicted to be generally less effective--in this case, those treatments involving reading-only response mode; (b) lower ability students would profit proportionally more than would higher ability students, from experimental treatments hypothesized to be generally more effective--in this case those treatments involving overt response mode. The composite rankings in figures 15 and 17 offer evidence to support these hypotheses.

The Long General Test provided three of the significant T x R x A interactions. Figure 18 gives an overview of the mean percent correct scores and the lines which show change over time for the experimental conditions embodied in the T x R x A interactions.

Figure 18 gives a graphic reason for the non-significant T x R x A interaction at immediate post-acquisition testing and at immediate post-review testing, as well as giving an equally graphic reason for the significant T x R x A interactions found between post-acquisition and delayed-retention, post-review and delayed-retention, and for the delayed-retention percent correct scores, all on the Long General Test.

Figure 18

Mean per cent correct on the Long General Test: text treatment by review treatment by ability level.



Other Analyses.

A number of analyses not central to the purpose of this research were performed. The analysis of variance tables may be found starting with Table XIV in Appendix B. Included there are analyses made of scores at the post-acquisition and the post review testing times.

Time to Complete the Instructional Program.

The analysis of time taken to complete the initial instructional programs indicated that significantly less time was spent by Ss on the Reading-Only program than on the Overt Response program ($p < .001$). An average of three hours was spent on the program by overt responding Ss for every two hours spent by reading-only Ss. This 3/2 ratio was accepted as demonstrating that substantially less time was spent in initial instruction by Reading-Only Ss.

This time difference is of particular interest, because the overt response program contained a smaller proportion of response-requests than the usual linear program. One might expect such a factor to minimize time differences. However, overt-responding Ss undoubtedly spent additional time not only in overt response (constructing answers on the paper) but also in the unmeasured covert activity of generating (or "thinking up") the response. For samples from each of the initial instructional programs see Appendix C.

Table 18 gives mean percent correct scores on each measuring instrument at the delayed-retention testing time. The table shows scores for students who had overt response mode versus those who had reading-only response mode on initial instruction.

Table 18

Mean percent correct score on measuring instruments at delayed retention testing, for students having had overt response versus reading-only response on initial instruction.

Measuring Instrument	Reading-Only Response Mode	Overt Response Mode
Short General	58.27	59.20
Long General	46.62	47.28
Chronology	40.30	40.66
Concepts	16.35	19.33
Sum	161.84	166.47
Mean	40.46	41.62

The differences in scores between the response modes were too small to be significant, but they become important in an inverse way when one realizes that half-again as much time was spent completing the program by overt responding students.

Teacher and Classroom Differences.

Because the experiment was analyzed with the six classroom replications combined into one, there was some curiosity on the author's part as to whether the effects of experimental treatments might have differed by some non-random factor other than students' ability from one of the original classroom groups to the other. For the satisfaction of this curiosity, medium ability students were selected for a special analysis. This level of ability was chosen because there were at least a few medium ability students in each of the six original classrooms, while this was not true for either the high or the low ability level. The analysis was made with the Short General Test on gain scores from pre-acquisition to delayed-retention testing. The results did not show a significant difference at or below the 5 percent level (df: 5/26; MS_{error}= .88 $MS_{classrooms}$= 1.30; F = 1.47; $p > .20$).

Ability and Different Measuring Instruments.

At the time of delayed-retention testing, the four measuring instruments were assessed for differential difficulty. The relative difficulty levels were consistent for the three ability groups. Mean scores are given in Table 19.

Table 19

Mean Percent Correct Scores on each Different Measuring Instrument, by each Ability Level--Long-Term Delayed-Retention Testing.

Ability Level	Measuring Instrument (test)			
	Short General	Long General	Chronology	Concepts
High	72.39	61.47	54.72	34.49
Medium	60.91	47.99	39.37	13.86
Low	41.67	29.06	25.84	4.59

In view of the hypotheses concerning interactions of ability level with experimental treatments, and in view of experimental results verifying the same, it was felt that the reader might be interested in whether or not there was some disproportional effect on any ability level from any measuring instrument. As one can see in Table 20, there was nothing but what might be normally expected owing to mental ability and measuring instruments of differential difficulty. This finding indicated that the variation within the instructional and review materials, rather than the various measuring instruments themselves, produced the effects shown by this experiment.

The measuring instrument of greatest difficulty was the Concepts Test, next were the Long General and the Chronology Tests; the instrument of least difficulty was the Short General Test.

Analysis of variance showed significant effects for both test type and ability level, but no interaction between the two. These results are shown in Table 20.

Table 20

Analysis of Variance: Mean Percent Correct Scores on each Different Measuring Instrument, by each Ability Level--Long-Term Delayed Retention Testing

Source	df	MS	F	p
Test	3	878.53	90.09	<.001*
Ability	2	929.04	97.38	<.001*
T x A	6	6.73		
ERROR	464	9.54		

*indicates effect accepted as reliable $p < .05$

Multiple-Choice and Open-End Test Formats.

Scores on the multiple-choice forms of the tests were substantially higher than were scores on the open-end forms of the tests ($p < .01$), thus substantiating the hypothesis that recognition would produce higher criterion test scores than would recall.

Gain Scores and Test Format. The measuring instruments at the time of delayed retention testing were composed of a random selection of both m-c and o-e items (See Appendix C). For this reason, only those items from the long-delayed test which were of the same format as was given to the particular S on his pre-acquisition or post-acquisition tests were used in obtaining the particular gain or loss scores used in analysis. Since the assignment of Ss for earlier testing times to m-c or o-e formats was made at random, and since the designation of which particular criterion test-item would be m-c or o-e at the delayed-retention testing time was also completely at random, no consistent bias should be reflected in the results obtained from this method of obtaining gain scores. (An analysis which compared gains for just the two formats--($m\text{-}c_2$ - $m\text{-}c_1$) versus ($o\text{-}e_2$ - $o\text{-}e_1$)--indicated no significant differences between them; $F < 1.00$.) That is, when only multiple-choice differences were subtracted from open-end differences, there were no significant differences in the "differences". On this basis, difference (gain or loss) scores for the two formats were combined whenever appropriate in analysis.

Reliability. Soon after the post-acquisition test was completed by all students a check was made on the test reliability of both the multiple-choice and the open-end forms of the Long General Test. Data were used from two classe which appeared to have the most consistently homogeneous grouping. One of the classes (Class 4) had consistently higher ability Ss, and the other (Class 5) was composed mostly of lower ability Ss. The coefficients of reliability, using the split-half method and computed by the Spearman-Brown modified formula were:

(4) high ability class, multiple-choice: .63
Open-end: .83
(5) low ability class, multiple-choice : .74
open-end : .88

These coefficients were sufficiently high that the reliability of both formats of the instrument was considered acceptable for both higher and lower ability Ss.

Discussion and Conclusions

Overt versus Reading-Only Response Mode.

The findings concerning response mode were, on the whole, quite different between the General tests and the Chronology Test. The two General tests indicated different effects from response mode for students of different levels of ability; such that higher ability students profited more from reading-only, while lower ability students were more favorably affected by overt response. However, when the Chronology Test was used to assess the long-term effects of the chronological review materials, a significant main effect was found for students of all abilities, in favor of overt responding during review. Although the differences were small we felt they were particularly interesting because they represented the only main effect in the entire experiment which was reflected consistently for students of all ability levels.

The problems created by the overt response question are much more basic and much broader than any problems which have to do with writing styles or programming styles, or whatever is the current cause célèbre. The fact that no consistently favorable effect for overt response was found for all students with general continuous discourse materials in this particular experiment should not necessarily shake anyone's convictions that overt response could have a beneficial effect for most students' learning from continuous discourse materials. The problems with this belief multiply as materials get further and further from the one-to-one associates level, and it becomes more and more difficult for the person preparing the instructional instrument to determine what kind of overt responding should be most facilitative. Whatever was accomplished in the Learning About Communism program, it obviously did not provide the most suitable overt response procedures for the acquisition of general knowledge, for students of all abilities. On the other hand, the chronological review materials which employed the overt response mode were effective for all students.

We accepted these findings as confirming the hypothesis that overt response would be much more likely to be beneficial for the chronological materials rather than for continuous discourse materials in this study.

At our present state of research competence, with what kinds of materials should we be able to demonstrate a substantial value for forms of practice more active than ordinary reading? The overt response technique used with chronological review materials in the present study was quite successful. Experiments with paired associates are relatively easy to quantify, and an optimum schedule of demonstration, practice, and even review might be established for any given audience and set of associates.

This ease of quantification should also hold true for serial-mechanical tasks, where the stimulus and response members are set up rather clearly in a more or less one-to-one relationship. There have been, in fact, a wide variety of experiments conducted with serial-mechanical tasks, the results of which might be offered as a starting point for serious consideration of the values of various kinds of overt responding.

Apparently, then, the more simple and direct the learning situation, the easier it will be to determine useful overt response procedures. With paired associates it is quite clear what must be practiced in order for the learner to reach criterion. It is equally clear what the criterion will be. If a program to teach English-Italian word equivalents is needed, and the students are English speaking, then the Italian words need to be learned and the connections between the known English and the newly-learned Italian words need to be practiced. There is not very much room for ambiguity as to what words are relevant for the learner to practice. Even more important, in terms of criterion performance, the test situation will be almost identical to the instructional situation. Such comments apply more or less to the learning of a procedural-mechanical or physical skill.

When we examine many learning situations, however, (especially school learning situations) this one-to-one relationship is much more difficult to find. Some learning tasks are direct--such as learning to spell or learning vocabulary equivalents, while other tasks are extremely indirect, such as generalizing from philosophy or history. One of the reasons often advanced for teaching history to children is that if they know what has happened in the past, they will be able to generalize to the possibilities in present day and even future situations. Whether or not this is a very tenable hypothesis, the materials which would be used to test it become immensely more complex and much more difficult to analyze.

Suppose that the criterion task in a history class is to apply what the student knows about Alexander the Great to the prediction of the kind of society he might create and maintain if he lived today. What words in an autoinstructional program about Alexander the Great for example, will have been relevant to write into blanks or to choose in a multiple-choice situation, in order to help the student attain the stated objective? We do not know the answer, but surely a vastly different looking instructional instrument would be required than the common product on the market today.

The closer the everyday criterion performance will resemble what the studen has been practicing during instruction, the easier it will be to determine relevant overt response procedures for the instructional situation. And with these closely corresponding instruction and performance situations, the experimenter in his turn should, at least according to the present theory, be able to demonstrate some advantage for the overt response procedures. The more distantly related the desired criterion behavior is to the instructional situation, the less advantage that will probably be demonstrable for overt response procedures in the program; possibly not because overt responding is not useful, but because it is so much harder to find out what kind of overt response procedures would be useful.

It appears that there are several questions which both the conscientious writer of instructional materials and the serious experimenter need to be able to answer. These questions are related to the overall problems of objectives specification and objectives realization:

First, if one is planning to construct an instructional instrument employing the generally more time-consuming devices of overt response, does he know what things need to be practiced and what connections need to be established? Second, can he afford to spend the time, money and technical effort which would really be necessary in order to find out what the better forms of practice might be? If not he might better spend his efforts in writing and testing a really good reading-only program.

Spaced versus Massed Practice.

Many investigators have attempted to define the influences of distributed practice in connection with meaningful learning; some have met with a certain amount of success, but most have not. (See Underwood, 1961; 1964 for a description of a large number of efforts directed more or less toward this goal, and for a brief discussion of his theory which evolved from those efforts.)

Although there were significant interactions involving the distribution of practice schedule employed in the present experiment, no significant main effects were found favoring either the spaced or the massed schedule.

Significant interactions of Practice, with some other variable or variables were found: 1. with the Long General Test for percent correct scores at the time of delayed-retention testing (T x R x P, showing some very slight overall favorability for massed practice excepting two conditions OT/OR and RT/RR, for which spaced practice was superior, and R x P x A, which indicated a general favorability for massed practice for all ability levels, excepting the medium ability level having overt response on review--for which spaced practice was superior); 2. with the Chronology Test for short-term effects of review (P x A, showing that medium ability students profited from spaced review); 3. with the Long General Test for long-term effects of review (T x R x P, showing spaced practice conditions to be neither the very best nor the very worst, and massed practice conditions to be either the very best--with a different response mode for review than for initial text--, or the very worst--with the same response mode for review as for initial text); and 4. with the Chronology Test for long-term effects of review (R x P, showing that massed practice with reading-only review was significantly poorer than any other condition and R x P x A, which indicated that if the student had massed practice he was a good deal better off with overt rather than reading-only response mode on review).

These interactions are extremely ambiguous; moreover, this variable was quite gross and very arbitrary in this study. The huge possibilities for devising quite different but equally arbitrary schedules for review indicate to the author that--no matter what the results had been--it would be difficult to defend the findings against speculations that vastly different results might have been found with only small changes in the schedule. And such different results could not be adequately explained by the present level of technology dealing with meaningful learning. Computerized monitoring of individual performance might indicate some, at least partly generalizable, patterns for review.

For our purposes it was concluded that an almost infinite set of schedules of demonstration and practice might have been and might still be invented and tested; but since we had neither the time nor the funds to do this, it has begun to appear that a new view of distribution of practice for meaningful material might be useful. Programed instruction is the relevant general model for the present research project, and in the next experiments, kinds of distributed practice will be considered which are basic to current theories about the autoinstructional materials themselves, for instance, theories about the spacing of practice which is provided by 'small steps' in an autoinstructional program.

The assumption might be made that proper programming of initial instructio or review should organize content and break it down into small or easily comprehensible steps. For that reason it is perhaps not as important to superimpose artificial spacing of practice in terms of the amount of time which passes between the making of each step. Short of the limits of sheer physical exhaustion of the student, perhaps the program itself should be expected to provide adequate spacing of practice or review. One way of testing this hypothesis would be to vary the density of the subject-matter content, both in number of repetitions of any single idea and in number of opportunities for overt or otherwise planned active practice--that is, make the steps more condensed and longer. This variation could then be viewed as 'massing' as opposed to the 'spacing' given in the original version of the program. This procedure will be used for the second set of experiments in this series.

The Relationship between Instructional Materials and Criterion Tests.

With each of the measuring instruments, the best or most sensitive discriminations between treatments were made among middle ability students. The next most sensitive discriminations were made among students of high ability, and the least sensitive discriminations appeared among students of low ability. In most instances the final assessments did not discriminate among low ability students at all, and because the final test scores for these students were so low, we have interpreted the results to be less a function of sensitivity of the measuring instruments than a function of how small an amount low ability students learned overall.

The General Tests.

The General tests are reasonable samples of the instructional content of the initial text. No differential content emphasis is reflected in these measuring instruments, because the items were drawn at random from a pool of general items covering the entire text content.

General materials of the nature involved in the Communism program are of the type most difficult for which to write or justify specific behavioral objectives, although paper and pencil tests over this kind of information can be produced with great facility. This fact alone must often lead us to deceive ourselves about the true objectives of such materials, since the tacit, implied behavioral objectives, in terms of school curriculum are usually not paper and pencil objectives at all. It remains to be seen whether more novel methods for assessing the effects of learning from such continuous discourse materials would be able to reflect larger overall differences between experimental treatments involving manipulation of variables within the instructional instrument.

The Chronology Test.

Chronological materials are the sort with which it is fairly simple to obtain precision and replicability in experimental practice--conditions which all too often do not obtain with complex subject matters programed in continuous discourse.

Although one cannot fully justify using the Chronology Test as a part of the normal classroom assessment situation, the results from experimental assessments with this test were particularly gratifying. The initial instructional text of course included the chronological information which appeared on the Chronology Test, but no specific attempt to teach this information by programming of overt responses was made until the time of review. This probably accounts not only for the small effect from either overt or reading-only response mode measured by the Chronology Test at the immediate post-acquisition testing time, but also for the positive effect which chronological review, especially overt review, had on long-term retention.

The Concepts Test.

Except for the main effect owing to Ss' ability, no effects were demonstrated with the Concepts Test. For this reason no discussion of specific experimental treatments can be made in connection with the Concepts Test.

The author would, however, like to indulge in a few speculative remarks about this measuring instrument. First of all, every ability level, including the high ability level, had low criterion scores and great loss scores on the Concepts Test, which indicated to us that a large amount of difficult material was either not made clear to the students or was not practiced sufficiently by them in the initial instruction and review materials. Therefore, a great deal too much was expected as criterion behavior for this test of conceptual knowledge. Second, since none of the variables employed in the experimental treatments indicated a positive effect on the Concepts Test for anyone (even for high ability students), it might be concluded that formation of concepts and understanding of difficult factual material would be better facilitated by means other than those employed in this experiment. On the other hand, it might just as easily be argued that the student population simply has not attained a developmental level or a maturity level which would enable them to grasp these concepts from any instructional system.

If one believes that the variables or the method of materials presentation used in this study ought to aid in the formation of concepts, then one might conclude that the variables and method would somehow have to be employed to greater advantage in order to produce a measurable effect on this particular criterion instrument.

Efficiency.

It was not our intention to make analyses on efficiency scores (criterion test score divided by time taken to complete the program), as such. There were a number of reasons for this decision. First, the author has previously encountered a logical difficulty with this type of data, where the range of scores in any one treatment group can be expected to be fairly large. A certain spirit of the equation (efficiency = score/time) is lost when a student who spent 1 minute and got a score of 1 has equal efficiency to the person who spent 60 minutes and got a score of 60. Second, decisions regarding efficiency (score, with time taken into account) over effectiveness ('pure score') or vice versa, often involve economic considerations or other value judgments, and are usually concerned with "size of differences"--a quantity not formally supplied by most methods of data analysis. In the present research, the overall time taken to complete each version (overt versus reading) of the program was analyzed This avoided most of the problems with efficiency scores, since a substantial time saving with little or no criterion-test-score loss was easily demonstrated.

The Influence of Student Ability.

In an experiment of the present type there are a number of problems in materials themselves, for example, what kind of behavioral objectives can be and should be specified, what presentation format and what level of discourse should be used, what kind of measuring instruments will be correct to assess the effects of the instructional materials and to measure how nearly the behavioral objectives have been attained. However, above and beyond the difficulties with materials, there are additional, and very real problems posed by the student population. These problems can be many and various, centering mostly around motivation to learn and level of ability. Under the heading of ability are other factors: reading speed and comprehension, general cultural sophistication, and attention span, to name several.

In the present study, it is fairly easy to see why one kind of response mode or distribution of practice might be more favorable for higher ability students and another kind might be more favorable for lower ability students. The higher ability student usually reads much faster than his lower ability classmate; he also understands much more of what he reads the first time through. Because of his ability advantage the higher ability student has been able to satisfy more of his curiosity about the world around him and, therefore, he often has greater cultural sophistication than his lower ability classmate, to apply in the learning of a new task. The higher ability student normally has a longer attention span than the lower ability student, which indicates that distributions and schedules of practice should be different for high and low ability students.

It makes sense, then, that the higher ability students in this experiment would have profited at least equally well and perhaps more from reading-only response mode on the more easily comprehensible continuous discourse materials than he would from making overt responses to small fixed-size steps. The lower ability student, on the other hand, was somewhat benefited by being led step by step through the overt-response program--but with reading-only materials he possibly became embroiled in complex knowledge and new vocabulary, and eventually lost interest.

Long-term Retention: Comparisons of this Study with the Results of Other Studies.

It is, perhaps, appropriate to end with a comparison of the results of this study with the results of some other investigations. A number of inferences can be made from this comparison--even if these cannot be tested directly as hypotheses.

Table 18 shows gains and losses for the present study, over a seven month period from post-acquisition to delayed-retention, and compares these with losses over the comparable length of time from three studies mentioned earlier in this paper (Pressey and Robinson, 1944, p. 543). Results from the Long General Test and the Chronology Test are shown. The Short General Test and the Concepts Test are excluded, since assessments over the same period of time were not made for those two tests.

Although the extent of favorability for the General review procedure used in our study could not be completely determined from the intra-experimental data, there is a strong indication, shown in the inter-experimental comparison below, that retention of general subject-matter was better in the present experiment than in the other experiments being compared. This could be attributed to better initial instruction, the effect of review instruction, or both. For results on the Long General Test, this author is inclined to favor the "both" hypothesis. However, since no deliberate attempt was made to teach chronological information in the initial instructional program, and since there was gain rather than loss for chronological information over the review period, the author is inclined to favor the "effect of review instruction" hypothesis, for results on the Chronology Test.

There is also a clear indication that over this time period the higher ability students lost less and gained more than did lower ability students. This effect may be observed in the 'Difference' column of Table 18.

Table 18

Comparison of Gains and Losses for Several Subject-Matters over Seven Months' Time

Ability	Immediate Post Acquisition Per Cent Correct	Delayed Retention Per Cent Correct	Difference
	- 7 months -		
	Learning About Communism		
	Long General Test		
High	64.68	61.47	- 3.21
Medium	52.75	47.49	- 4.76
Low	37.33	29.06	- 8.27
Mean Total	51.59	46.17	- 5.42
	Chronological Test		
High	46.82	54.72	+ 7.90
Medium	38.22	39.37	+ 1.15
Low	24.96	25.84	+ .88
Mean Total	36.67	39.94	+ 3.31
	*High School Chemistry Test		
Mean Total	65.00	53.00	-12.00
	*Beginning Psychology Test		
Mean Total	77.00	44.00	-33.00
	*Grade School History Test		
Mean Total	70.00	57.00	-13.00

*This information comes from the graph on page 543 in Pressey and Robinson (1944).

References

Adams, J. A. & Reynolds, B. Effect of shift in distribution of practice conditions following interpolated rest, *J. exp. Psychol.*, 1954, *47*, 32-36.

Alter, Millicent. Retention as a function of length of retention interval, intelligence, and training time. *J. Prog. Inst.*, 1963, *2*, 7-17.

Alter, Millicent & Silverman, R. E. The response in programed instruction. *J. Prog. Inst.*, 1962, *1*, No. 1, 55-78.

Bassett, S. J. John Hopkins University Studies in Education, 1928, No. 12. In S. L. Pressey & F. P. Robinson, *Psychology and the new educator*. New York: Harper Bros., 1944.

Briggs, L. J. & Angell, D. Programed instruction in science and mathematics. *Rev. educ. Res.*, 1964, *34*, 354-373.

Briggs, L. J. & Hamilton, Nancy R. Meaningful learning and retention: Practice and feedback variables. *Rev. educ. Res.*, 1964, *34*, 545-559.

Brown, D. R., Naylor, J. C., & Briggs, G. E. The retention of discrete and continuous tasks as a function of the interim practice with modified task requirements. *USAF AMRL - TDR*, 1963, No. 63-35.

Eigen, L. D. Technology and educational practice. Paper presented at the American Psychological Association Convention. In *Background portfolio*. New York: The Center for Programed Instruction, 1961. 11 pp.

Evans, J. L., Glaser, R., & Homme, L. E. A preliminary investigation of variation in the properties of verbal learning sequences of the "Teaching Machine" type. In A. A. Lumsdaine & R. Glaser, *Teaching machines and programed learning*. National Education Association of U. S., Washington, D. C.: 1960, 446-451.

Evans, J. L., Glaser, R., & Homme, L. E. An investigation of "teaching machine" variables using learning programs in symbolic logic. *J. educ. Res.*, 1962, *55*, 433-452.

Goldbeck, R. A. & Campbell, V. N. The effects of response mode and response difficulty on programed learning. *J. educ. Psychol.*, 1962, *53*, 110-118.

Green, E. B. The retention of information learned in college courses. In S. L. Pressey & F. P. Robinson, *Psychology and the new educator*. New York: Harper Bros., 1944.

Hamilton, Nancy R. Effects of logical versus random sequencing of items in an autoinstructional program under two conditions of covert response. *J. educ. Psychol.*, 1964, *55*, 258-266.

Hamilton, Nancy R., Willis, Mary B., Bivens, L. W., & Doolin, D. J. *Learning about Communism*. Palo Alto, Calif.: American Institutes for Research, 1963.

Holland, J. G. *Design and use of a teaching machine*. Paper presented to American Psychological Association Symposium, 1960. Cambridge: Harvard Univer., 1960 (mimeo).

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. *Experiments on Mass Communications*. Princeton: Princeton University Press, 1949.

Johnson, I. O. *J. educ. Psychol.*, 1930, 21:37-47. In S. L. Pressey & F. P. Robinson, *Psychology and the new educator*. New York: Harper Bros., 1944, p. 543.

Jones, H. E. Experimental studies in college teaching. *Arch. Psychol.*. 1923, No. 68.

Keislar, E. R. & McNeil, J. D. A comparison of two response modes in an autoinstructional program with children in the primary grades. *J. educ. Psychol.*, 1962, *53*, 127-131.

Krumboltz, J. D. & Weisman, R. G. The effect of overt versus covert responding to programed instruction on immediate and delayed retention. *J. educ. Psychol.*, 1962, *53*, 9-92.

Lambert, P., Miller, D. M., & Wiley, D. E. Experimental folklore and experimentation: The study of programed learning in the Wauwatosa Public Schools. *J. educ. Psychol.*, 1962, *55*, 485-494.

Landauer, T. K. Two states of paired-associate learning. *Psychol. Rep.*, 1962, *11*, 387-389.

Lindquist, E. F. *Design and analysis of experiments in psychology and education*. Cambridge; Houghton Mifflin, 1953.

Luh, C. W. The conditions of retention. *Psych. Monogr.*, 1922, Vol. 31, No. 142.

Lumsdaine, A. A. (Ed.) *Student response in programmed instruction: A symposium on experimental studies of cue and response factors in group and individual learning from instructional media*. Washington, D. C.: National Academy of Sciences--National Research Council, 1961.

Madsen, M. C. Distribution of practice and level of intelligence. *Psychol. Rep.*, 1963, *13*, 39-42.

McGeoch, J. A. *The psychology of learning: An Introduction*. New York: Longmans, Green and Co., 1946.

McNemar, Q. *Psychological Statistics*. New York: John Wiley and Sons, 1962.

Naylor, J. C. & Briggs, G. E. Long-term retention of learned skills: A review of the literature. ASD TR 61-390, Ohio State Univer., Columbus, Ohio, 1961.

Naylor, J. C., Briggs, G. E., & Reed, W. G. The effects of task organization, training time, and retention interval on the retention of skill. USAF AMRL - TDR, 1962, No. 62-107.

Powers, S. R. In S. L. Pressey & F. P. Robinson Psychology and the new educator. New York: Harper Bros., 1944, p. 543.

Pressey, S. L. A machine for automatic teaching of drill material. School and Society, 1927, 25, No. 645. In A. A. Lumsdaine & R. Glaser (Eds.) Teaching machines and programmed learning: A source book. Washington, D. C.: National Education Association, 1960, 42-46.

Pressey, S. L. & Robinson, F. P. Psychology and the new educator. New York: Harper Bros., 1944.

Shettel, H. H., Clapp, D. J., & Klaus, D. J. The application of a by-pass technique to programmed instruction for managerial training. Pittsburgh, Pennsylvania: American Institutes for Research, July, 1963. AIR-C74-7/63-FR.

Sones, A. M. & Stroud, J. B. Review, with special reference to temporal position. J. educ. Psychol., 1940, Vol. 31, 665-676.

Spitzer, H. F. Studies in retention. J. educ. Psychol., 1939, Vol. 30, 641-656.

Stolurow, L. M. & Walker, C. A comparison of overt and covert response in programed learning. J. educ. Res., 1962, 55, 421-429.

Underwood, B. J. Ten years of massed practice on distributed practice. Psychol. Rev., 1961, 68, 229-247.

Underwood, B. J. Forgetting. Scientific American, 1964, 210, (3), 91-99.

Williams, Joanna P. Comparison of several response modes in a review program. J. educ. Psychol., 1963, 54, 253-260.

Wittrock, M. C. Verbal stimuli in concept formation: learning by discovery. J. educ. Psychol., 1963, 54, No. 4, 183-190.

Appendix A

Description of the Experimental Population

Appendix A

Description of the Experimental Population

Table i

Statistics Descriptive of the Population
Total SCAT Scores (percentiles) for Each Class in the Experiment

Class	Teacher	Mean	Mode	Median	Standard Deviation
1	I	80.65	95	87.5	19.14
2		74.50	95	79.5	22.01
3	II	64.08	85	70.5	54.46
4		94.73	95	93.94	5.83
5	III	50.94	75	53.79	22.35
6		71.31	95	78.5	22.12

Table ii

Correlation Coefficients (r)
The Long General Test correlated with various Population Measures,
for Each Class and for All Classes in the Experiment

	STEP math		STEP reading		SCAT total		SCAT verbal		SCAT quantitive	
Class	o-e	m-c	o-e	m-c	o-e	m-c	o-e	m-c	o-e	m-c
1	.08	.74	.38	.54	.41	.66	.27	.66	.48	.66
2	.75	.66	.81	.71	.82	.57	.80	.65	.74	.38
3	.22	.27	.57	.62	.57	.69	.58	.67	.02	.55
4	-.10	.12	.26	.25	.77	.59	.77	.12	.56	.71
5	.90	.62	.64	.54	.83	.55	.69	.52	.87	.51
6	.63	.43	.29	.78	.78	.72	.77	.66	.62	.70
All Classes Together	.65	.64	.56	.70	.76	.73	.72	.67	.67	.62

Appendix B

Analysis of Variance Tables

Appendix B

Analysis of Variance Tables

Gain over the Entire Experiment.

Table I

Analysis of Variance indicating effects measured by the Short General Test, from pre-acquisition to delayed-retention.

Source	df	MS	F	p
Text	1	.00		
Review	1	.00		
Practice	1	.24		
Ability	2	5.92	7.79	<.001*
T x R	1	.06		
T x P	1	.54		
T x A	2	.80	1.05	
R x P	1	.06		
R x A	2	.68		
P x A	2	.68		
TxRxP	1	.15		
TxRxA	2	2.99	3.93	<.025*
TxPxA	2	.64		
RxPxA	2	.13		
TxRxPxA	2	1.94	2.55	<.10
ERROR	95	.76		

*Indicates an effect accepted as reliable (p <.05)

Effects at Delayed-Retention Testing.

Table II

Analysis of Variance indicating effects measured by the Long General Test, at delayed-retention.

Source	df	MS	F	p
Text	1	2.61		
Review	1	32.57		
Practice	1	16.73		
Ability	2	2358.16	38.64	<.001*
T x R	1	90.56	1.41	
T x P	1	130.01	2.03	<.20
T x A	2	6.68		
R x P	1	38.41		
R x A	2	6.24		
P x A	2	135.44	2.12	<.20
TxRxP	1	717.77	11.21	<.005*
TxRxA	2	357.26	5.58	<.01*
TxPxA	2	38.36		
RxPxA	2	244.97	3.83	<.05*
TxRxPxA	2	.00		
ERROR	95	64.01		

*Indicates an effect accepted as reliable (p <.05)

Table III

Analysis of Variance indicating effects measured by the Short General Test, at delayed-retention.

Source	df	MS	F	p
Text	1	5.19		
Review	1	1.16		
Practice	1	7.13		
Ability	2	1896.52	62.80	<.001*
T x R	1	10.14		
T x P	1	22.58		
T x A	2	.83		
R x P	1	18.17		
R x A	2	62.42	2.07	<.20
P x A	2	27.92		
TxRxP	1	24.36		
TxRxA	2	66.88	2.21	<.20
TxPxA	2	9.98		
RxPxA	2	46.71		
TxRxPxA	2	3.64		
ERROR	95	30.20		

*Indicates effect accepted as reliable ($p < .05$).

Table IV

Analysis of Variance indicating effects measured by the Chronology Test, at delayed-retention.

Source	df	MS	F	p
Text	1	.78		
Review	1	.15		
Practice	1	29.96		
Ability	2	1749.62	43.08	<.001*
T x R	1	3.79		
T x P	1	121.23	2.98	<.10
T x A	2	5.41		
R x P	1	19.22		
R x A	2	31.27		
P x A	2	104.80	2.58	<.10
TxRxP	1	99.64	2.45	<.20
TxRxA	2	81.58	2.01	<.20
TxPxA	2	8.56		
RxPxA	2	98.63	2.43	<.10
TxRxPxA	2	116.87	2.88	<.10
ERROR	95	40.61		

*Indicates effect accepted as reliable ($p < .05$).

Table V

Analysis of Variance indicating effects measured by the Concepts Test, at delayed-retention.

Source	df	MS	F	p
Text	1	5.28		
Review	1	43.08		
Practice	1	.24		
Ability	2	2006.52	45.72	<.001*
T x R	1	1.50		
T x P	1	43.56		
T x A	2	86.20	1.96	<.20
R x P	1	2.76		
R x A	2	85.68	1.95	<.20
P x A	2	69.64	1.59	
TxRxP	1	3.29		
TxRxA	2	57.30	1.30	
TxPxA	2	26.68		
RxPxA	2	78.97	1.80	
TxRxPxA	2	5.47		
ERROR	95	43.89		

*Indicates effect accepted as reliable ($p < .05$).

Short-term Overall Effects.

Pre-Acquisition to Post-Review with the Short General Test.

Table VI

Analysis of Variance indicating effects measured by the Short General Test from pre-Aquisition to post-review.

Source	df	MS	F	p
Text	1	.20		
Review	1	.34		
Practice	1	.01		
Ability	2	16.88	8.07	<.001*
T x R	1	5.12	2.45	<.20
T x P	1	.18		
T x A	2	4.18	2.00	<.20
R x P	1	2.02		
R x A	2	4.47	2.14	<.20
P x A	2	3.95	1.89	<.20
TxRxP	1	5.60	2.67	<.20
TxRxA	2	3.10	1.48	
TxPxA	2	.75		
RxPxA	2	2.63	1.26	
TxRxPxA	2	.19		
ERROR	117	2.09		

*Indicates effect accepted as reliable ($p < .05$).

Effects of Review.

Short-Term Effects of Review.

Table VII

Analysis of Variance indicating effects measured by the Long General Test, from post-acquisition to post-review.

Source	df	MS	F	p
Text	1	4.23		
Review	1	10.45		
Practice	1	.41		
Ability	2	17.96	3.87	<.05*
T x R	1	2.65		
T x P	1	4.70		
T x A	2	4.51		
R x P	1	.00		
R x A	2	4.67	1.01	
P x A	2	4.38		
TxRxP	1	.54		
TxRxA	2	1.38		
TxPxA	2	2.58		
RxPxA	2	5.10	1.10	
TxRxPxA	2	1.22		
ERROR	112	4.64		

*Indicates effect accepted as reliable ($p < .05$).

Table VIII

Analysis of Variance indicating effects measured by the Chronology Test, from post-acquisition to post-review.

Source	df	MS	F	p
Text	1	.65		
Review	1	.96	1.31	
Practice	1	.91	1.24	
Ability	2	4.42	6.05	<.005*
T x R	1	.03		
T x P	1	.74	1.01	
T x A	2	.10		
R x P	1	.09		
R x A	2	1.75	2.39	<.10
P x A	2	4.29	5.87	<.005*
TxRxP	1	.43		
TxRxA	2	1.30	1.78	<.20
TxPxA	2	2.06	2.82	<.10
RxPxA	2	1.83	2.50	<.10
TxRxPxA	2	.44		
ERROR	90	.73		

*Indicates an effect accepted as reliable ($p < .05$).

Long-Term Effects of Review.

Table IX

Analysis of Variance indicating effects measured by the Long General Test, from post-acquisition to delayed-retention.

Source	df	MS	F	p
Text	1	3.72		
Review	1	.00		
Practice	1	.36		
Ability	2	9.76	2.44	<.10
T x R	1	5.76	1.44	
T x P	1	.07		
T x A	2	1.28		
R x P	1	.00		
R x A	2	1.76		
P x A	2	1.38		
TxRxP	1	16.74	4.18	<.05*
TxRxA	2	15.25	3.81	<.05*
TxPxA	2	1.70		
RxPxA	2	8.95	2.24	<.20
TxRxPxA	2	1.38		
ERROR	94	4.00		

*Indicates effect accepted as reliable (p <.05).

Table X

Analysis of Variance indicating effects measured by the Chronology Test, from post-acquisition to delayed-retention.

Source	df	MS	F	p
Text	1	.01		
Review	1	5.88	7.95	<.01*
Practice	1	2.23	3.01	<.10
Ability	2	1.14	1.50	
T x R	1	.02		
T x P	1	1.08	1.46	
T x A	2	.80	1.08	
R x P	1	3.11	4.20	<.05*
R x A	2	.33		
P x A	2	2.04	2.76	<.10
TxRxP	1	.04		
TxRxA	2	3.74	5.05	<.01*
TxPxA	2	.56		
RxPxA	2	2.80	3.78	<.05*
TxRxPxA	2	.47		
ERROR	92	.74		

*Indicates effect accepted as reliable (p <.05).

Effects Owing to the Passage of Time.

Table XI

Analysis of Variance indicating effects measured by the Long General Test, for the passage of time from post-review to delayed-retention.

Source	df	MS	F	p
Text	1	.02		
Review	1	.38		
Practice	1	1.22		
Ability	2	4.18	1.61	
T x R	1	13.41	5.18	<.05*
T x P	1	.00		
T x A	2	.52		
R x P	1	.00		
R x A	2	1.20		
P x A	2	1.20		
TxRxP	1	1.90		
TxRxA	2	10.58	4.08	<.025*
TxPxA	2	.30		
RxPxA	2	5.84	2.25	<.20
TxRxPxA	2	3.02	1.17	
ERROR	91	2.59		

*Indicates effect accepted as reliable (p <.05).

Table XII

Analysis of Variance indicating effects measured by the Chronology Test, for the passage of time from post-review to delayed-retention.

Source	df	MS	F	p
Text	1	.05		
Review	1	.00		
Practice	1	1.56	2.42	<.20
Ability	2	1.08	1.68	<.20
T x R	1	.00		
T x P	1	.22		
T x A	2	.67	1.04	
R x P	1	1.56	2.42	<.20
R x A	2	.16		
P x A	2	1.96	3.04	<.10
TxRxP	1	1.26	1.96	<.20
TxRxA	2	.34		
TxPxA	2	.80		
RxPxA	2	1.90	2.95	<.10
TxRxPxA	2	.16		
ERROR	92	.64		

*Indicates effect accepted as reliable (p <.05).

Table XIII

Analysis of Variance indicating effects measured by the Concepts Test, for the passage of time from post-review to delayed-retention.

Source	df	MS	F	p
Text	1	200.40	2.54	<.20
Review	1	104.40	1.32	
Practice	1	45.36		
Ability	2	240.56	3.05	<.10
T x R	1	27.00		
T x P	1	21.72		
T x A	2	120.70	1.53	
R x P	1	299.46	3.79	<.10
R x A	2	119.00	1.51	
P x A	2	100.24	1.27	
TxRxP	1	126.63	1.60	
TxRxA	2	47.56		
TxPxA	2	89.98	1.14	
RxPxA	2	234.12	2.97	<.10
TxRxPxA	2	94.79	1.20	
ERROR	95	78.96		

*Indicates effect accepted as reliable (p <.05).

Miscellaneous Analysis.

Table XIV

Analysis of Variance indicating effects measured by the Long General Test, post-acquisition.

Source	df	MS	F	p
Text	1	.72		
Form	1	626.40	205.41	<.001*
Ability	2	247.98	79.74	<.001*
T x F	1	1.44		
T x A	2	6.64	2.14	<.20
F x A	2	6.01	1.93	<.20
TxFxA	2	.45		
ERROR	154†	3.11		

*Indicates effect accepted as reliable (p <.05).
†Accounts for all Ss, including no-review Ss, since this measurement was made before review treatments.

Table XV

Analysis of Variance indicating effects measured by the Chronology Test, post-acquisition.

Source	df	MS	F	p
Text	1	.24		
Form	1	48.00	120.88	<.001*
Ability	2	7.96	20.04	<.001*
T x F	1	.07		
T x A	2	.16		
F x A	2	.02		
TxFxA	2	.11		
ERROR	154†	.40		

*Indicates effect accepted as reliable (p <.05).

†Accounts for all *S*s, including no-review *S*s, since this measurement was made before review treatments.

Table XVI

Analysis of Variance indicating effects measured by the Concepts Test, post-review.

Source	df	MS	F	p
Text	1	1.50		
Review	1	25.96	3.18	<.10
Practice	1	7.80		
Ability	2	578.02	70.91	<.001*
T x R	1	.80		
T x P	1	10.77	1.32	
T x A	2	.90		
R x P	1	.08		
R x A	2	5.94		
P x A	2	12.92	1.58	
TxRxP	1	.46		
TxRxA	2	13.24	1.62	
TxPxA	2	4.44		
RxPxA	2	4.94		
TxRxPxA	2	3.82		
ERROR	114	8.15		

*Indicates an effect accepted as reliable (p <.05)

Appendix C

Samples from the Instructional Texts and the Measuring Instruments

Appendix C

Samples from Instructional Texts.

Overt Response

66. Throughout the years, since Lenin first interpreted the theories of Marx, the Communists have been using various methods to try to gain control of other countries. Three of these methods are: 1) direct conquest, 2) revolution, and 3) election.

67. At the end of each of the following will you write whether it is an example of gaining control by 1. direct conquest, 2. revolution, or 3. election:

Suppose Communist China invaded India again and took over a large piece of Indian territory.

Direct conquest

(Which method of gaining control?)

Reading-Only Response

66. Throughout the years, since Lenin first interpreted the theories of Marx, the Communists have been using various methods to try to gain control of other countries. Three of these methods are: 1) direct conquest, 2) revolution, and 3) election.

67. The following examples illustrate these various ways of gaining control:

   1. Suppose Communist China invaded India again and took over a large piece of Indian territory.

This is an example of direct conquest.

Overt Response

68. Suppose rioting broke out in Guatemala and the government was overthrown. When a new government was set up, it decided that it wanted to be like the government of Cuba and get help from Soviet Russia.

revolution
(Which method of gaining control?)

1. direct conquest
2. revolution
3. election

---

69. The Soviet Union has absorbed several of her small neighbors, such as Latvia, Lithuania, and Estonia. In the same way Communist China conquered Tibet.

This method for gaining control of a country is called Direct conquest
(Which method of gaining control?)

1. Direct conquest
2. Revolution
3. Election

---

70. Suppose economic conditions got worse in Argentina and the people insisted that a new election be held under United Nations supervision. In the election the Communist Party got more votes than any other party.

Election
(Which method of gaining control?)

1. Direct conquest
2. Revolution
3. Election

Reading-Only Response

68. 2. Suppose rioting broke out in Guatemala and the government was overthrown. When a new government was set up, it decided that it wanted to be like the government of Cuba and get help from Soviet Russia.

This is an example of gaining control by revolution.

---

69. 3. The Soviet Union has absorbed several of her small neighbors, such as Latvia, Lithuania, and Estonia. In the same way Communist China conquered Tibet.

This is an example of gaining control by direct conquest.

---

70. Suppose economic conditions got worse in Argentina and the people insisted that a new election be held under United Nations supervision. In the election the Communist Party got more votes than any other party.
This is an example of gaining control by election.

Overt Response

71. Which two of the three methods of gaining control of a country require some cooperation from the people living within that country?

Revolution and Election.

Which one of the three methods of gaining control may be used without the cooperation of the people in that country? Direct conquest requires no cooperation.

---

72. Which of the three methods of gaining control of a country has been least successfully used by the Communists? Election

---

73. Name the three methods that Communists might use to gain control of a country:

direct conquest

revolution

election

---

74. Once the Communists have gained control of a country, they install a Communist government. This government uses various methods to keep the people in the newly-won country under control.

Reading-Only Response

71. Of the three methods of gaining control of a country, two require some cooperation from the people living within that country. Those are the methods of revolution and election.

The method of gaining control which may be used without the cooperation of the people in that country is direct conquest.

---

72. The method of election has been least successfully used by the Communists to gain control of a country.

---

73. In summary, the three methods that Communists might use to gain control of a country are:

1. Direct conquest 2. Revolution 3. Election

74. Once the Communists have gained control of a country, they install a Communist government. This government uses various methods to keep the people in the newly-won country under control.

Overt Response

267. We have seen that the conflict between the Democratic and Communist ideologies has resulted in a divided world. The goal of the Communist nations is to dominate the world and convert it into a Communist society. This goal is opposed by the nations of the Free World, who feel that each country should be free to elect its own form of government.

---

268. In pursuing their goal of world domination, the Communists have often resorted to force. Communist nations have started several small-scale or "brushfire" wars in order to gain more and more territory. The free world often finds it necessary to use force, armies, etc. to oppose invasion of countries that do not want a Communist government.

Reading-Only Response

267. We have seen that the conflict between the Democratic and Communist ideologies has resulted in a divided world. The goal of the Communist nations is to dominate the world and convert it into a Communist society. This goal is opposed by the nations of the Free World, who feel that each country should be free to elect its own form of government.

---

268. In pursuing their goal of world domination, the Communists have often resorted to force. Communist nations have started several small-scale or "brushfire" wars in order to gain more and more territory. The free world often finds it necessary to use force to oppose invasion of countries that do not want a Communist government.

Overt Response

269. In 1950, Communist North Korea invaded the Republic of South Korea. To drive the Communists out, free nations which were a part of the organization called the United Nations sent an army of free-world troops, most of them United States soldiers.

---

270. When the U. N. armies drove the North Korean forces back to the border of China Chinese Communists joined the fight and forced the U. N. troops back into South Korea. Despite this intervention by the Communist Chinese army, U. N. troops were (were/ were not) able to prevent North Korea from gaining any South Korean Territory.

Reading-Only Response

269. In 1950, Communist North Korea invaded the Republic of South Korea. To drive the Communists out, free nations which were a part of the organization called the United Nations sent an army of free-world troops, most of them United States soldiers.

---

270. When the U. N. armies drove the North Korean forces back to the border of China, Chinese Communists joined the fight and forced the U. N. troops back into South Korea. Despite this intervention by the Communist Chinese army, U. N. troops were able to prevent North Korea from gaining any South Korean territory.

Overt Response

324. The Communist Party has been the subject of many restrictions and legal investigations in the United States, because many responsible people do not think the United States Communists are keeping good faith with the definition of a legal political party. That is, they think that the United States Communists may be cooperating with foreign governments in attempting to overthrow our system of government by force.

325. We do not (do/do not) want any group of people to be able to change our form of government without our consent.

326. In controlling activities against the government, the Communist countries can prevent their citizens from getting together to discuss issues and make plans about anything the Communists disapprove of. The free countries generally allow their people to organize into groups which believe differently from the government in power. Therefore, the free nations often have to deal with organized Communist activities, while the Communist countries do not permit (permit/ do not permit) any organized anti-Communist activity.

Reading-Only Response

324. The Communist Party has been the subject of many restrictions and legal investigations in the United States, because many responsible people do not think the United States Communists are keeping good faith with the definition of a legal political party. That is, they think that the United States Communists may be cooperating with foreign governments in attempting to overthrow our system of government by force.

325. We do not want any group of people to be able to change our form of government without our consent.

326. In controlling activities against the government, the Communist countries can prevent their citizens from getting together to discuss issues and make plans about anything the Communists disapprove of. The free countries generally allow their people to organize into groups which believe differently from the government in power. Therefore, the free nations often have to deal with organized Communist activities, while the Communist countries do not permit any organized anti-Communist activity.

Overt Response

327. The United States Constitution guarantees our citizens freedom of speech and the right to assemble peacefully for any purpose which is not illegal. Would it be legal in our country for one political party to form groups of police to prevent people from getting together to discuss issues or from voting for the candidates of other parties on election day?

no.

---

328. In Communist countries a person who is suspected of wanting some system of government other than Communism may be arrested without a warrant; he may be held in jail indefinitely without a trial; and if he should get a trial, it may not be a trial by jury.

What would happen to our system of government if we decided for example, we would put all the people who are not members of the Republican Party in jail? It would not be a democracy. It would be a dictatorship.

Reading-Only Response

327. The United States Constitution guarantees our citizens freedom of speech and the right to assemble peacefully for any purpose which is not illegal. It would not be legal in our country for one political party to form groups of police to prevent people from getting together to discuss issues or from voting for the candidates of other parties on election day.

---

328. In Communist countries a person who is suspected of wanting some system of government other than Communism may be arrested without a warrant; he may be held in jail indefinitely without a trial; and if he should get a trial, it may not be a trial by jury.

If we decided, for example, to put all the people who are not members of the Republican Party in jail, our system of government would cease to be a democracy and would be a dictatorship.

Appendix C

Measuring Instruments.

The delayed retention testing was made with this instrument which includes the questions from every test. On test forms which originally had both multiple-choice or open-end formats, half of the questions were randomly selected to be included at delayed-retention testing in multiple-choice format; the other half appear in open-end format.

PRE on page 1, before question 1, indicates the set of questions from the Short General Test.

RCH on page 4, before question 1, indicates the set of test questions from the Russian Chronology.

CCH on page 3, before question 1, indicates the set of test questions from the Chinese Chronology.

I&D on page 8, before question 2, indicates the set of questions from the Concepts Test.

Questions 1, 2, and 3 on page 1 may be used for the purpose of illustrating what was meant by "parallel open-end and multiple-choice formats". The parallel form of question 1 was: The initials of the name of the strongest free world alliance are: NATO. The parallel form of question 2 was: Pick the reasons (there may be more than one) why outlawing the Communist Party in the United States is not the best way to deal with the threat of Communism: a + c. a. The most important Communist activity is illegal and underground. b. The Communists would disappear from America. c. Our counterspies can sometimes join a legal Communist Party in order to find out its illegal activities. d. Americans would forget the threat of Communism. The parallel form of question 3 was: A person is most likely to be a Communist if he believes: d a. That Americans want to dominate the world. b. That Russia will attack the U. S. c. Everyone in Russia should own some land. d. That the State should have complete power over the individual.

Your Name ______________________

Your 8th Grade Teacher and Class ______________________

Write the letter, or letters, of your answer in the blank which follows a multiple choice question. Or write your answer to "fill-in" questions.

PRE

1. The initials of the name of the strongest free world alliance are: b

a. SEATO	c. UNICEF
b. NATO	d. ECSC

2. Outlawing the Communist Party in the United States is not the best way to deal with the threat of Communism because: (1) The most important Communist activity is illegal and secret. (2) Our counterspies can sometimes join a legal Communist Party in order to find out its illegal activities.

3. A person is most likely to be a communist if he believes that the State should have complete power over the individual.

4. In the Soviet Union, how many out of each hundred people in the population are Communist party members? five.

5. One of the main differences between Soviet and American justice is: c
   - a. All Russian prisoners are sent to concentration camps.
   - b. There are better lawyers in America, but better laws in Russia.
   - c. In Russia a person may be arrested without a warrant, be held in prison indefinitely without a trial, and will not be given a trial by jury.
   - d. In America everybody but the Communist Party members has Constitutional rights.

7. The Soviet Constitution provides that people can vote in elections by: c.
   - a. Open ballot.
   - b. Indirect ballot.
   - c. Secret ballot.

Your Name________________________

Your 8th Grade Teacher and Class________________________

8. What is the main difference between the attitude and work habits of the American farmer who runs his own farm, as compared to the Russian farmer who works on a collectivized farm? b.

a. The Russian farmer works harder because his produce goes to the State.
b. The American farmer produces more because he owns his own land.
c. The Russian farmer is lazier than the American farmer.
d. The Russian farmer can produce more because he is not limited by production quotas like the American farmer.

9a. If an American citizen wanted to join the Communist Party: c

a. He would be shot as a traitor.
b. He would be deported to Russia.
c. Our laws would permit him to join.
d. He would be put in jail.

9b. If a citizen of a Communist country wanted to join an anti-Communist party: c.

a. He would be allowed to join.
b. He would be sent to America.
c. He would not be allowed to join.
d. He would have to register.

10. Which large Communist country has the greatest natural resources? The Soviet Union.

11. Chiang Kai-shek was driven to the island of: b, in the year d.

a. Sicily	a. 1901
b. Formosa	b. 1917
c. Cuba	c. 1940
d. Japan	d. 1948

12. Collectivization of agriculture was forced on the Russian peasants for three reasons; give one of these reasons: individual ownership was against communist beliefs.

Your Name____________________

Your 8th Grade Teacher and Class____________________

13. In what year did Lenin die? 1924 .

14. Sun Yat-sen's revolutionary group was know as the Nationalists. The revolution in China was over in 1911 (date).

15. The Chinese Communist leader who led the campaign to drive Chiang Kai-shek out of China is Mao Tse-tung .

16. Two countries the Communists have tried to take over by force are a and c .

a. Korea
b. Japan
c. Vietnam
d. Turkey

17. The nuclear test ban treaty recently signed by the United States and the Soviet Union forbids exploding nuclear devices in at least three specific places. One of these is outer space. What are the other two?

underwater

in the atmosphere

18. What was the nature of the relationship Marx saw between a country's political-economic system and its social-moral structure? a .

a. That the political-economic system determines the social-moral structure of a country.

b. That politics are more important than the moral structure.

c. That social and moral structure determines the political and economic structure of a country.

d. A country's morals are reduced by politics.

19. While Marx was editing a radical socialist newspaper in Germany in c he met b .

a. 1917
b. 1905
c. 1842
d. 1852

a. Lenin
b. Engles
c. Tito
d. Mao Tse-tung

Your Name________________________

Your 8th Grade Teacher and Class________________________

20. How is a "class society" different from the communist society which Marx had in mind? A communist society is not supposed to group people according to their money, power, or prestige.

RCH

1. Marx and Engels a. Trotsky and Engels
b. Lenin and Marx
c. Marx and Engels wrote the
d. Lenin and Trotsky

Communist Manifesto in c a. 1818
b. 1842
c. 1848
d. 1867

2. What three major events in Russian history occurred in 1917? d e h
   a. The Social Revolutionary Party was organized.
   b. The New Economic Policy was established.
   c. Lenin was exiled to Siberia.
   d. The Tsarist regime was overthrown by the Social Revolutionary Party.
   e. Lenin returned to Russia and became leader of the Bolshevik Revolution.
   f. World War 1 ended.
   g. The Communists were victorious in the Russian Civil War.
   h. The Bolshevik Revolution seized power from the provisional government.
   i. Lenin returned to Russia and organized the Russian Social-Democratic Party.

3. The capital of Russia in 1917 was Petrograd, .

4. The New Ecomomic Policy (NEP) was established by Lenin (whom) and was applied for 7 years and 1921 (date) until 1928 (date).

Your Name____________________

Your 8th Grade Teacher and Class____________________

5. Lenin died in __c__ a. 1921 and __a__ a. Stalin
b. 1923 b. Trotsky
c. 1924 c. Krushchev
d. 1928 d. Bulganin

became ruler of Russia. He inaugurated the first __c__
a. New Economic c. Five-Year
b. Farm Collectivization d. Seven-Year

Plan in __b__ a. 1924 c. 1927
b. 1928 d. 1929

6. Russia became allied with the United States in World War II after __Hitler__ (who) broke his alliance agreement with Russia in __1941__ (date).

7. In __d__ a. 1939 __b__ a. Mao Tse-Tung
b. 1941 b. Tito
c. 1945
d. 1948

broke with Stalin and made __b__ a. Mainland China
b. Yugoslavia an

independent Communist state.

8. Stalin died in __d__ a. 1945 c. 1951
b. 1949 d. 1953

9. By __1958__ (date), Khrushchev was the undisputed dictator of the Soviet Union.

CCH

1. After the revolution in China in __1911__ (date), Sun Yat-sen made China a __Republic__ (what form of government).

Your Name______________________

Your 8th Grade Teacher and Class______________________

2. After Dr. Sun Yat-sen died in __b__ a. 1921 c. 1927
b. 1925 d. 1931

__b__ a. Mao Tse-tung
b. Chiang Kai-shek secured appointment as Commander of
the __c__ a. Choomintang c. Kuomintang Army
b. Communist Army d. Army of the Central Peoples Republic
and began using the army to unify the country.

3. During the years of war between Japan and China, the __Communists__ and the __Nationalists__ cooperated in a fairly unsuccessful effort to drive the Japanese out of China.

4. The Chinese fought the Japanese from __1937__ to __1945__
a. 1931-1936 c. 1937-1941
b. 1937-1945 d. 1935-1942 The Chinese were hard pressed at the time the United States entered the war against Japan in __b__.
a. 1939 c. 1942
b. 1941 d. 1945

5. __Chiang Kai-Shek__ (who) was elected President of China in __1948__ (date).

6. The __b__ a. Nationalists in China seized __(not scored)__
b. Communists
a. Mongolia c. Malaya and obtained a large amount of war
b. Manchuria d. Formosa
supplies which had been left behind by the Japanese after World War II.
With these supplies they made war on the __b__ a. Communists
b. Nationalists
c. Japanese
The __b__ a. Nationalist Party was successfully in
b. Chinese Communist
power in China in __c__ a. 1946 c. 1949
b. 1948 d. 1954

Your Name____________________

Your 8th Grade Teacher and Class____________________

7. What three main events occurred in 1949 to make that year a significant date in Chinese history? b e f

   a. Chou En-lai became Premier of China.

   b. Chiang Kai-shek resigned as President of the Chinese republic.

   c. Chiang Kai-shek was elected President of the Chinese republic.

   d. Mao Tse-tung resigned as Chairman of the Peoples Republic.

   e. Chiang Kai-shek went to Formosa (Taiwan).

   f. Mao Tse-tung became Chairman of the Central Peoples Republic.

8. The current Premier of the Peoples Republic of China is Chou En-lai
The most powerful man in the Chinese government is Mao Tse-tung
who is Chairman of the Communist Party

I&D

2. Lenin's most important contribution to Communist philosophy was the idea that the leaders in a revolution must be: b .

   a. Members of the proletariat

   b. Dedicated professionals

   c. Members of a minority group.

   d. Peasants.

3. Both Russia and Communist China believe that the rest of the world will finally become Communist. What is the difference between the Russian idea and the Chinese idea of how this will come about? Russians believe in economic competition - Chinese believe in direct conquest

4. What is the most important reason why there are some legal restrictions placed on the Communist Party in the United States? c

   a. Because we do not give Constitutional rights to people we don't like.

   b. Because the Communist Party is good, as long as it has legal restrictions.

   c. Because some responsible people feel that the Communists may attempt to overthrow our government by force.

   d. Because the Communist Party does not have sincere members.

Your Name____________________

Your 8th Grade Teacher and Class____________________

5. Which two of the following reasons are the purposes of the first Five-Year Plan in Russia? b c

   a. To conserve natural resources.
   b. To bring about rapid industrialization of Russia.
   c. To collectivize agriculture.
   d. To strengthen foreign relations.

7. How many Communist party members are there in Indonesia? d

   a. 3,000 c. 250,000
   b. 1,350,000 d. 2,000,000

8. At the present time, United States industry is ~~more~~ producing more than Soviet industry.

9. As an economic system capitalism is characterized by competition and by the fact that businesses operate to make a profit. A third important characteristic of capitalism is private ownership.

11a. When was the Chinese Communist Party founded: b.

   a. 1917 c. 1911
   b. 1921 d. 1925

b. Who was one of the principal founders of the Chinese Communist Party? c

   a. Chiang Kai-shek c. Mao Tse-tung
   b. Khrushchev d. Dr. Sun Yat-sen

12. Some countries are pro-Communist and some are anti-Communist; other countries are uncommitted.

   Albania is Pro-Communist.
   (what political affiliation?)

   Turkey is anti-Communist
   (what political affiliation?)

   India is Uncommitted.
   (what political affiliation?)

13. In d a. 1941 c. 1949
b. 1945 d. 1950, the Chinese Nationalist government
in b a. Mainland China c. Japan
b. Formosa d. Korea began to receive substantial
economic and military aid from: c a. Russia c. United States
b. Japan d. Taiwan

14. In area, how many times larger than the United States are Mainland China and the Soviet Union together? d
a. 4 c. 2½
b. 2 d. 3½

15. Which large, uncommitted nation has had territorial disputes with Communist China? India.

17. What is the difference between the goals of the Communist nations and the nations of the Free World? Communist nations want to make the whole world communist.

19. The Chinese Communist leader who led the campaign to drive the Nationalists out of China is: a
a. Mao Tse-tung c. Sun Yat-sen
b. Chiang Kai-shek d. Chou En-lai

20. Kuomintang means Nationalist party in Chinese.

21. What fraction of all the people in the world are controlled by the Communists? 1/3. Are all these people Communist Party members? (not scored)

DR
2. Describe the dialectical principle by filling in the blanks below:

The thesis is opposed by the antithesis, dausing a revolution which results in a synthesis, which becomes a new thesis opposed by the antithesis causing another revolution which results in a new synthesi. Marx thought that the last perfect synthesis would be communism.

Your Name____________________

Your 8th Grade Teacher and Class____________________

3a. The three methods Communists use to gain control of other countries are:

1) direct conquest
2) revolution
3) election

4a. List three economic advantages the Russian people have compared to Americans:

1) prices for bread
2) income taxes
3) security of employment

4b. List four economic advantages the American people have compared to the Russians:

1) housing
2) prices for clothing
3) wages for skilled workmen
4) meat prices

6. If a person is arrested in the United States, he can get a writ of habeas corpus **which requires that the authorities** bring a prisoner into court to determine legality of keeping him in jail.

7b. The original ideals of the founders of the Communist movement have been corrupted by the Communist doctrine that "the end justifies the means**."**

AMERICAN INSTITUTE FOR RESEARCH

mailing address: Station A, P.O. Box 11487, Palo Alto, Calif.
Telephone 321-2130 (Area Code 415)
office location: 1791 Arastradero Road, Palo Alto, California

AIR-D49-7/65-FR-II

INCREASING LONG TERM RETENTION OF KNOWLEDGE

Experiment II: Instruction from Small Step versus Large Step Programs with Immediate versus Delayed Review Procedures for Students of Different Ability Levels

Nancy Russell Hamilton

Barbara D. Porteus

Grant Number Title VII: 7-48-7670-204
National Defense Education Act of 1958
Leslie J. Briggs, Principal Investigator

Final Report Experiment II
July 1965

The Research Reported Herein Was
Supported by a Grant from the
U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education

Abstract

Portions of three published autoinstructioanl programs--one on basic instruction in decimals arithmetic, one on squaring two-digit numbers ending in five, and one on the Geography of China--were administered at different times during the school year, all three programs to each of three classes of sixth grade students. Each program took from one-half to three hours of student time.

The experimental design was 2 x 2 factorial. (1) Review and redundant materials were separated from the original instructional versions and were given separately at a later date to one-half of the students. This variation was designated Integrated- versus Delayed-Review. (2) The original mainline of instruction, which was left after removal of redundancy and review, was condensed into larger, even less redundant steps, with fewer overt response requests than in the original instructional version. The larger steps were given to one-half of the students. This variation was designated Small- versus Large-Steps. The four resulting treatments were: (1) Integrated-Review/Small Steps (the program as originally published). (2) Delayed-Review/Small Steps (the program as originally published, with review and redundancy removed to a later date). (3) Integrated-Review/Large-Steps (review and redundancy inserted within the Large-Steps mainline of instruction, in as nearly as possible the same places as it occurred in the original published version). (4) Delayed-Review/Large-Steps (the Large-Steps mainline of instruction, with review and redundancy removed to a later date).

Data were analyzed with test scores for all three classes pooled, for each of the three experiments, and according to three levels of ability (high, medium, and low, based on ranges of scores on the School and College Ability Test). Results indicated that the individual program was the source of the greatest variation in results, and that level of ability had the expected effect in terms of higher ability students learning and gaining more in general. An analysis was made of immediate effects of total instruction (control for recency), using gains from pre-acquisition to post-acquisition testing for Integrated-Review students and from pre-acquisition to post-review testing for Delayed-Review students. On this analysis both the Decimals and Geography of China programs showed significantly more gain on test scores for students who had the small step versions of the mainline of instruction. For overall, long-term effects from pre-acquisition testing to delayed-retention testing 94 days later, there were significant effects owing to level of ability, for all three programs. Only the Decimals program, however, still showed the significant main effect in favor of small steps on the original mainline of instruction.

No statistically-significant effects were found for the Integrated- versus Delayed-Review variable, for either immediate or long-term gains. Materials in conditions which had small steps in the mainline of instruction took significantly longer for students to complete.

These results should be generalized only to other instructional instruments and experimental conditions very similar to those employed in the present study.

ACKNOWLEDGEMENTS

The research described in this report was made possible through the intelligent and enthusiastic cooperation of the Las Lomitas School District in Menlo Park, California. District personnel who were exceptionally helpful to us were Mr. Lawrence J. Fuller, Superintendent of Schools; Mrs. Margaret Giffin, Curriculum Specialist; Mr. William Carrico, Principal, La Loma School; Mr. Richard Jensen, Principal, Las Lomitas School, Mr. Wilmer Larsen, Principa Ladera School, Mr. Edward Loveless, Principal, La Entrada School, and Miss Jacque Hamilton, Sixth-Grade Teacher at Ladera School, who provided us with her most welcome and useful specialized knowledge during a large part of one summer vacation.

The participating teachers were Miss Jacque Hamilton, Sixth Grade, Ladera School, Mr. Gary Kilarr, Sixth Grade, Ladera School, and Mr. Thomas Rhoads, Sixth Grade, Las Lomitas School. Without the assistance and cooperation of these teachers this set of studies could not have been conducted.

Assisting during experimental sessions were Miss Patti Morse, Mrs. Alice Frimmersdorf, and Mrs. Virginia Sturwold. Invaluable aid was given by Miss Sally Richerson and Miss Brenda Bordwell in preparation and typing of manuscripts.

We would like to express our gratitude to Dr. Edwin B. Parker, of the Communications Department at Stanford University, and to Mr. David Markle, of American Institutes for Research for reading and commenting on the final draft of the report; and to all the staff members at American Institutes for Researc who have provided us with their assistance and moral support.

Table of Contents
Experiment II

Page

Problem. - 1

Background. - 1

Review of Literature. - 3

Size of Step.- 3

Scheduling of Review.- - - - - - - - - - - - - - - - - - - 4

Method.- 6

Experimental and Procedural Designs.- - - - - - - - - - - - - - 6

Subjects. - 8

Materials.- 8

Instructional Programs.- - - - - - - - - - - - - - - - - - 8

Criterion Tests. - 10

Procedure.- 10

Rationale and Hypotheses.- 11

Small Versus Large Steps. - - - - - - - - - - - - - - - - - - - 11

Integrated versus Delayed Review. - - - - - - - - - - - - - - - 13

Ability.- 13

Time to Complete Instructional Materials. - - - - - - - - - - - 13

Results. - 14

Analysis of the Data. - 14

Results Overall.- 14

Overall Effects in Terms of Time Spent on Instruction.- - - - - - 22

Immediate Effects of Total Instruction: Control for Recency. - - 22

Additional Analyses.- 26

Table of Contents
Experiment II (continued)

Page
Discussion and Conclusions. - 26

References. - 29

Appendix A. - 31

Appendix B. - 56

INCREASING LONG-TERM RETENTION OF KNOWLEDGE

Experiment II: Instruction from Small Step versus Large Step Autoinstructional Programs with Immediate versus Delayed Review Procedures for Students of Different Ability Levels

Problem

These experiments were the final set in a group designed to use autoinstructional programs for both initial instruction and review and to assess the effects of this on long term retention of knowledge. The specific purpose of this second set of experiments as a whole was to investigate two different strategies for determining optimal amounts and distribution of demonstration, practice and review for meaningful continuous discourse materials. The first strategy involved varying the density of instructional materials, by varying both the content redundancy and the amount of overt response in each frame. ("Frame" is the name given to the individual instructional unit in an autoinstructional program.) We named this independent variable "size-of-step". The second strategy involved varying the time at which review was administered--integrated with the initial instructional material or delayed until several weeks following initial instruction--we called this independent variable "review".

The reason for conducting a set of experiments, rather than a single experiment, was to compare and contrast criterion test performances when the same students and the same independent variables were employed with various different programed subject-matters. Three published autoinstructional programs were used, one on elementary decimals arithmetic, one on the geography of China, and one teaching a method for squaring two-digit numbers ending in five.

Background.

Many fairly consistent indications for successful schedules of practice have been provided by investigations using paired-associates of low meaningfulness. However, when attempts are made to assess the results of distributed practice using materials with a high degree of meaningfulness, very few investigators have been able to obtain conclusive evidence favoring either massed or spaced practice.

The findings from the earliest experiment in the present project (Experiment I) were also inconclusive on the distribution of practice dimension. In Experiment I length of and spacing of practice sessions during review were varied. It was tentatively concluded from this experiment that such "spacing or massing" of practice may be regarded as a gross construct, for which the crucial factor in obtaining significant effects would probably be student exhaustion. The author believes that the purpose of most experiments with schedules of practice is not to measure physical exhaustion of the student. Neither is the primary purpose, necessarily, to find schedules of demonstration and practice which are worse than those commonly employed in everyday situations. Rather the purpose of such investigations is to find schedules of demonstration and practice which facilitate learning beyond those commonly employed in everyday situations.

This set of experiments was designed primarily to investigate the notion that retention in school subjects probably depends largely upon the quality of initial instruction. Other theories concerning retention were not a formal part of this investigation.

In these studies, two methods for varying 'review' were investigated. The first varied the amount of redundancy (review) in the mainline instructional sequence, and the second varied the time at which the actual review materials were presented: integrated with the initial instruction or delayed until a later time.

We began with the assumption that programing of either initial instruction or review materials should organize content and break it down into small or easily comprehensible steps. The theoretical constructs behind programed instruction and the rigorous development process which should be used in constructing a program should produce an instructional instrument which internally provides adequate spacing of practice or review (not to imply anything about external factors such as student exhaustion, nor to discuss "review" as normally scheduled for conventional instruction). Hence we took the unusual view that size-of-step in the program could be the major internal factor which determines spacing and amount of practice. According to this view, spacing of practice could be controlled by the amounts of redundancy and less crucial discourse which appear between new and more important facts or concepts

Underwood (1961; 1964) has developed a theory which attributes failure to retain information, to interfering associations. His theory is cogent and seems to offer various practical directions for validation and further extention through experimentation. In one article (1964) Underwood discussed his own research with schedules of practice within and between lists of words. He also mentioned research by Geoffrey Kappel of the University of California, Berkeley, that indicated favorable effects for gross spacings of time between learning sessions. His students, however, learned lists of several different sets of words paired to the same key words, and this is a task for which a great deal of proactive inhibition has previously been demonstrated. Worthwhile as these findings are, great caution should undoubtedly be exercised in attempts to use theory, which was developed with specific kinds of instructional materials, as an explanation for or as a basis for prediction of comparable effects from all sorts of instructional materials. For even if the principles of a theory are true, they may be very difficult to apply when learning conditions and instructional instruments are not sufficiently similar to those used in the original theoretical development. Research with continuous meaningful discourse is difficult and frustrating. Concrete behavioral objectives are hard to state and often difficult to justify. It is often very hard to define the kind of instructional situation which is represented by a specific set of materials so that sensible comparisons may be made to findings from some other specific set of materials.

Looking at interference from the purely stimulus side of the learning paradigm, then, one may argue that interference is a constant, for any individual, and thus that the initial learning is of more crucial importance than review learning. Looking at the learning paradigm from the response side, of course, the argument is much more complex, beginning with determination of how differently the individual might interact with his later environment as a result of his initial learning.

Underwood has indicated a need for the use of more meaningful verbal tasks in studies involving distributed practice. In the present study the size-of-step variation is internal to the instructional content. Varying the content and context density in continuous discourse materials has some similarity to the methodology employed by such investigators as Underwood. Findings with this size-of-step variable are then possibly generalizable to those from other investigations of distributed practice. The authors do not feel, however, that the timing-of-review variable in the present study is sufficiently similar to Underwood's conditions to lead us to expect closely comparable findings.

Review of Literature.

Size of Step.

Some investigations have not found significant differences in learning between small-step and large-step programs (e.g., Smith and Moore, 1961). A number of studies have, however, found such differences.

It should be noted that 'size-of-step' means different things to different people, e.g., strength of prompting of the learner's response, degree of steepness of student effort required, rapidity of new concept presentation, difficulty of the response, etc. Schramm (1962) compiled an annotated bibliography of research on programed instruction, which included research on size of step. He concluded that size of step, as a variable, had not been satisfactorily defined, but he noted nevertheless that when findings were significant the favorable effect on learning was generally from short steps.

Using college psychology Ss and a psychology program, Coulson and Silberman (1960) found that small step forms of programs (104 frames) produced a significantly higher post-test scores than did large step forms (56 frames). They also found that small step programs took significantly more time.

Evans, Glaser, and Homme (1960) used more and less condensed (larger and smaller steps) versions of a program on converting to number bases other than 10. They found that students who had been given small-step programs (51 and 68 frames) made significantly fewer program errors and performed significantly better on immediate and delayed retention tests than students with larger step programs (30 and 40 frames). Subjects were graduate students in education. A non-significant trend showed that the group with a 51-step program performed slightly better than the group with the 68-step program. This finding hints, at least, that there may be a point after which smaller and smaller step size would not facilitate learning.

Scheduling of Review.

One major reason for continued interest in grosser schedules of review lies in the school situation. The schools are not often concerned with, and in any case are not suitably equipped for experimentation with minute variations within instructional instruments. They are, however, concerned with factors which may critically effect the timing of basic instructional and review materials. Suppose for example, a fixed instructional core and set of review materials is available. Several interesting questions for school curriculum planning might occur: Is there any advantage to presenting the total core and review sequence to the student during a single time space (perhaps the first three weeks in a semester)? Or is there some advantage to giving the instructional core early and some additional review materials later in the semester? Or doesn't it matter which? If it doesn't matter, then curriculum planning can be accomplished with additional flexibility.

The notion that review must be scheduled as "review"--that is, sometime after initial instruction--is eminently sensible in the situation where a time-lapse occurs between the training program and the beginning of the job. In this case, review is provided as a refresher immediately before duties are assumed.

The school situation is somewhat different: The student is supposed to learn all or a portion of a subject in a specific length of time, say four and one half months. Usually, after this period of time, he progresses upward in his learning activities and out of the influence of the primary instructional situation. But in most instances what the student does next in school is assumed to be based partly on what he has learned before. Therefore, in a sense, review is built into the scale of progress. Because of this, the school situation, ever changing and enduring over a period of many years, presents retention problems of a different order from the practical training situation in which an immediate and repetitive kind of job performance is required.

It may be possible that whatever review is used could be integrated with the core instruction with no undesirable loss of long-term retention. In that case the complete instructional package could be scheduled at the time of initial presentation. It could, then, be concluded that it is not the review as such but rather the total amount of practice provided which has an influence on retention.

Following this argument for school training the problem of review would be regarded as identical to the problem of providing the best initial instruction. Findings from Experiment I in the present project certainly do not make this an untenable consideration, although there is evidence suggesting that giving additional emphasis to certain information (names, dates, and places, for instance) is beneficial. It is doubtful that such additional emphasis must necessarily be delayed until sometime after the initial instruction period.

The literature reviewed for Experiment I indicated, in general, that the amount and quality of initial instruction and practice was far more important in influencing retention than were any of the grosser schedules of review which were employed (e.g., Naylor et al., 1962). Also, there were strong indications that with sufficient practice at the time of initial instruction, the benefits which one normally expects to find from review provided at a later date do not always materialize (e.g., Brown, Naylor, and Briggs, 1963).

Such considerations may be lost in a discussion solely of independent variables. For instance, findings for the immediate-retention or post-review retention testing times may be misleading predictors of long-term retention of knowledge. In relation to this, if you believe that a subject in school--such as basic decimals arithmetic--has important short-term goals, then you have considerable interest in research which assesses short-term effects of instruction. If, on the other hand, you believe that the only important goals from a course are long-term, then you will have an interest in demonstrations that basic and primary instruction somehow determine whether the student will do well in later work. The intention of the present research lies somewhere between these extremes.

Gagne (1965) tells us that there are many kinds of learning appearing to have only short-term goals but which actually contribute to the student's performance of more sophisticated tasks later in his learning career--even though the origianl task may be largely forgotten.

It appears to the author that practice over a long period of time with some tasks (learning to tell time, for example) allows us to develop "bridges", or larger and larger steps, which we use to cross over or condense the more primary steps which we took in learning to perform the task. In other words, we do not have to keep prompting ourselves as heavily in order to continue to perform the task as we did when we were learning to perform it, and although we no longer take the smaller steps, we have not necessarily forgotten them. For instance, many of us learned to tell time by saying to ourselves, "the big hand is on six and the little hand is halfway between five and six; so it is one-half hour past five." We seriously doubt if there are many normal adults who, when asked the time, look at their watches and recite to themselves: "the big hand is on six..." On the other hand, we have not necessarily forgotten that this was how we learned to tell time--we teach our children to prompt themselves in just that way.

Most such discussions are without substantial experimental justification. However, they contain interesting notions, and they indicate that the best of all possible basic primary instruction should be provided. This is especially the case with school learning, since the student will not be formally reviewing the most basic subject-matters later on in his life.

Some of the literature on 'vanishing of cues and prompts' could contribute support to this notion (see, for example; Angell and Lumsdaine, 1962).

The problem of timing review for meaningful instructional materials has always been very difficult to investigate, since there are so many different scheduling possibilities for meaningful learning. Previous studies with programmed instruction have provided some useful information about scheduling of review. Glaser and Reynolds (1962) obtained favorable effects from a spaced practice condition where students learned two programed topics in sequence and subsequently repeated the first topic as review. Ferster (1960) also studied the effect of repeating a whole lesson in programed instruction. His subjects were medical school students, and the program was one on basic psychology (Holland and Skinner, 1961). He found no reliable differences in favor of repeating the lesson and concluded that any important repetition was already built into the program. However, for graduate students learning from the Holland and Skinner (1961) psychology program, Holland and Porter (1961) used a type of review in which students repeated the program items which they had initially answered incorrectly. It was found that this procedure led to fewer errors on the criterion tests both at the post-acquisition and six-month retention levels. Lumsdaine, Sulzer, and Kopstein (1961) found that overlay animation devices and additional examples increased learning from a film teaching micrometer usage to Air Force Basic trainees. They also found a "saturation" point, beyond which additional examples no longer facilitated learning. This saturation point was reached more quickly for Ss of lower intelligence (i.e., higher ability students continued to gain after lower ability students had reached a plateau), suggesting to those authors that mere repetition was not sufficient to improve performance.

Gagne (1965) suggested that for concept formation a variety of examples is needed to help the learner identify the important elements and discriminate them from other, less important elements in the instructional material.

Method

This rather unusual positioning of the Method section is necessitated by the fact that the experimental hypotheses were predicated partly on particulars of the methods and materials employed in this study.

Experimental and Procedural Designs.

The design paradigm used for the experiment conducted with each of the three programs is given in Figure 1 below. The delay between initial instruction and administration of review to students in delayed-review conditions and post-review testing was 17 days, plus or minus a weekend in some instances. Also for students in delayed-review conditions, the delay following review was 74 days, plus or minus a weekend in some instances. The total time for all Ss, from immediate post-acquisition testing to delayed-retention testing, for each experiment was 91 days.

Figure 1

Experimental design with indication of expectations for each experimental treatment at the delayed-retention measurement.

Step Size	Schedule of Review	
	Instruction with Review Integrated	Instruction with Review Delayed
Small-Step Programs	I. Hypothesized best condition (++), at delayed-retention measurement.	II. Hypothesized first intermediate condition (+), at delayed-retention measurement.
Large-Step Programs	III. Hypothesized second intermediate condition (-), at delayed-retention measurement.	IV. Hypothesized poorest condition (--), at delayed-retention measurement.

The procedural design and schedule are given in Figure 2, below.

Figure 2

Procedural design and schedule, including testing.

Portion of Treatment and Schedule of Tests	Program		
	Decimals	Squaring	Chinese Geography
Pre-acquisition test; initial instruction; post-acquisition test for all Ss.	October 8-9, 1964	November 2, 1964	November 24-25, 1964
Review for separate review conditions. A "post-review" test was administered to all Ss at this time.	October 26-27, 1964	November 19, 1964	December 11-14, 1964
Delayed-retention test. Administered to all Ss in a classroom at a single time.	January 8, 1965	February 1, 1965	February 26-March 1, 1965

Subjects.

A total of eighty students in three sixth-grade classes served as Ss in the experiments. Scores on the School and College Ability Test were available in school records for most students, and Ss in each class were matched into groups of four, according to similarity of mental ability as shown by their scores on the SCAT. From each ability-matched group, the four students were then randomly assigned, one to each of the four experimental treatments. Students for whom no ability scores were available in the school records were not included in the initial matching and assignment process. They were assigned to treatments at random and were assigned to the proper ability group for analysis if their ability scores became available soon enough for inclusion. Six Ss were excluded from the analyses (i.e., never became a part of the formal experiment) since either their intelligence test scores did not become available or data was incomplete owing to their withdrawing from school before the end of the experiment. On the 'Squaring' and 'Chinese Geography' programs there was no further S mortality. However, in the analyses of the 'Decimals' program, 3 high-ability Ss were excluded, because their pre-acquisition scores were greater than 95 percent.

Materials.

Instructional Programs.

The instructional materials were variations of the three previously-mentioned programs: Decimals and Per Cents (Smith, 1964); China: A Programmed Unit in Geography (Soens, et al., 1963); and Squaring Two-Digit Numbers Ending in Five (Angell and Campeau, 1962). The program (Learning About Communism, Hamilton, et al., 1963) used with eighth-grade students in Experiment I was not used in Experiment II, because it was too difficult for sixth-grade students. The experimental paradigm is reproduced below, indicating in each cell the kinds of materials the Ss received in each treatment condition.

Figure 3

Experimental paradigm showing kind of experimental materials used in each treatment condition.

Step Size	Schedule of Review	
	Instruction with Review Integrated	Instruction with Review Delayed
Small-Step (Original version of the program)	Original published program intact.	Mainline of published program as initial instruction with review delayed until 17 days after initial instruction.
Large-Step (Condensed version of the program)	Condensed mainline program intact. Review items in as nearly as possible the same places as in original published version.	Condensed mainline program, with review delayed until 17 days after initial instruction.

The desired variations in instructional instruments were produced without either adding to or subtracting from the basic minimal instructional content. Condensed or "large-step" versions of programs were produced by condensing the textual prose and omitting redundant information. Review programs were produced by taking review material intact from the original versions of programs.

The system followed for producing the various experimental materials was devised with the intention of its being as replicable as possible. The system worked as follows: 1. Each program "frame" was put on a separate card; 2. All items clearly labeled 'review' or 'remedial' in the original program were removed from the mainline of instruction and placed in the review-frame container (a box); 3. The following scheme was devised for making decisions as to which of the frames remaining in the mainline program would be called "review", for our experimental purposes: The first time an item of information, a fact, a notion, etc., was introduced by the original program, it was labeled D, meaning demonstration; the second time the same notion occurred it was labeled P, meaning practice; the third and every subsequent time the notion occurred, it was labeled R, meaning review, and was removed to the review-frame container. This completely arbitrary scheme worked surprisingly well for most of the information in the programs; however, there were instances where a notion was only used twice in the original program. When this occurred, the first instance was labeled D and P, meaning demonstration and practice, and the second instance was removed to review.

With only minor differences in individual judgments, this system could be used by anyone who cared to do so. The value of it or the validity of this method as a way of finding the review content of a program is not demonstrated in this research, and it would be up to the individual to decide whether or not to use it.

This method produced programed 'review materials' which were constant within each program for each experimental treatment, although review was of course, separate for some students.

Table 1 compares the three programs in terms of numbers of frames in each, and in terms of how these frames were finally distributed in the experimental treatments. Only the initial portions of the Decimals program and the Geography of China program were used, since we wanted to keep the maximum student time around two hours. The entire Squaring program was used, since it is quite short.

Number of frames is often a misleading way to judge the length or difficulty of a program; time taken to complete the program is a much better basis for judgment. These time data are given in the results section.

Table 1

Number of frames in each program and condition; mean time to completion.

Program and Step Size		Integrated Review		Delayed Review	
		Number of Frames		Number of Frames	
		Initial	Review	Initial	Review
Decimals Program	SS	209	0	76	133
	LS	156	0	23	133
Squaring Program	SS	54	0	32	22
	LS	42	0	20	22
Chinese Geography Program	SS	196	0	91	105
	LS	143	0	38	105

Criterion Tests.

In all, four tests were used for each subject-matter: a pre-acquisition test, an immediate post-acquisition test, a post-review test, and a delayed-retention test. In all three subject-matters the delayed-retention test was longest, and the other three tests were shorter and were sampled at random from the delayed-retention test.

These criterion tests were composed of items supplied as a part of the testing for the published program and of items constructed by the project staff and the teacher-consultant who worked with us. For each program a 44- to 88-item criterion test was constructed. The length of this test was dependent upon the nature of the subject-matter, and the longest test was for Geography of China, while the shortest test was for Squaring Two-Digit Numbers Ending in Five. These tests were given as the measures of delayed retention. The three shorter tests were sampled at random from the criterion test for each program, for the purpose of keeping track of the students' progress throughout the experiment (the pre-acquisition test, the post-acquisition test, and the post-review test). It was thought that comparable but non-identical criterion measures would be provided by sampling the shorter interim tests from the longer final test. This procedure was followed in order to minimize the possibility of test-items interacting with instructional items and thus confounding the instructional process with the testing process. (The delayed-retention measuring instrument for each program is given in Appendix B.)

Procedure.

The pre-acquisition test was given to all students in a classroom immediately before they began the initial instruction treatments. As a student finished his particular set of materials he gave them to the monitor.

Students were told that they weren't expected to know the answers to any or all of the questions on pre-acquisition tests. It was explained that some students might know the answers to questions, but no one was to worry over the test or try to guess answers to questions he didn't know. They were told that they would be given a chance on later tests to show what had been learned from the lessons. The student then completed the post-acquisition test, gave this to the monitor, and went to another room for other school work. At the time scheduled for review materials to be administered to delayed-review students, "post-review" tests were given at once to the integrated-review students; review materials and, later, the post-review tests were given to the delayed-review students. Therefore, students who had review integrated with initial instruction took the "post-review" test at the time students who had delayed review were completing review materials. The integrated-review students were then taken to a different room for other school work, while the delayed review students completed review treatments and took post-review tests. Students who had integrated review, therefore, had only the post-review test at the time when students who had delayed review were having both the review materials and the post-review testing. Each classroom was given the delayed-retention test in a single group.

Every student was given time to finish all tests and all instructional materials. We had two students who regularly worked so slowly that an extra hour or two was sometimes required for them to finish. For slow finishers a small quiet room and a staff monitor were provided at the school. Of the two extremely slow workers, one seemed to profit from the extra time spent, but the other performed irregularly on criterion tests--doing very well with Decimals, very poorly with Squaring, and intermediately well with Chinese Geography.

Rationale and Hypotheses

Small Versus Large Steps.

The literature on providing the learner with additional practice and additional examples indicated in several instances that there is a point beyond which additional practice or added examples will not improve performance. Basing our decision partly upon such findings, we decided for experimental purposes to accept the mainline of instruction in the original published version of each experimental program as having sufficiently small steps, enough examples, and adequate review materials, and we concentrated on composing variations in materials which would increase the size of the step and reduce the number of examples. In achieving this alternate form of the program, instruction was separated from the review (as described in the Materials Section of this paper), and the redundancy and amount of overt response was then varied in the instructional mainline so that the large-step version of the mainline of the program would have increased frame size, decreased redundancy, and fewer overt response requests. The actual review materials were unaltered for all treatment conditions, as the other independent variable in the experiments, "time of review", would have become unduly complicated by condensing the review materials as well as the mainline materials for large-step programs.

It should be expected that smaller steps in an autoinstructional program would facilitate learning and retention. It should also be expected that the smallness of step would not be merely smallness of context or smallness of stimulus. The predictions should of course, be contingent upon the notion that small content steps, or small mental steps, or smallness in terms of the response side of the programed instruction paradigm would facilitate learning more than simply a small context step or a small number of words in the stimulus. For example, consider the following two illustrations of step size:

> stimulus portion: 1) the "t"-test is a form of the analysis of variance; as such it is parametric, and the assumptions concerning the analysis of variance should be met when employing it. The "t"-test is a ____________________ test.
> (parametric/non-parametric)
>
> response portion: /parametric/

> stimulus side: 2) Obtain the "t" ratio between the following columns of figures:
>
> 2 5
> 3 6
> 4 6
> 5 8
>
> t = ____________
>
> response portion: /t = 3.22/

One would not really want to predict which of the above frames would produce greater learning, but there is no question as to which is the larger step--in terms of size of the mental step which the student must make before making the required overt response of writing an answer in the blank. The frame context or the stimulus portion of the first illustration is medium long, but the mental step which the student must make in order to obtain the answer is extremely small. Whereas, in the second illustration, the context is medium short, but the mental step(s) which the student must make to obtain the answer are many and complex.

For the present set of experiments the specific prediction was made that the small-step versions of the programs (the steps as originally published) would produce highest student criterion test and gain performance. It was further predicted that this effect would be most pronounced for the Chinese Geography program, since that program was judged by the staff and the participating teachers to contain "harder" subject-matter than the other two programs.

Integrated versus Delayed Review.

The method of obtaining review materials for the programs used in this set of experiments may be found in the Materials Section of this report. It can be seen that from the application of this method the resulting review sections were not materials in addition to the original sequence but rather were set aside from the original versions of the instructional programs to be used later. It was assumed for these experiments that the amount of redundancy and the portions labeled "review" in the published program had been tested and deemed necessary by the programmer and publisher in order for the student to attain his expected performance level. It was in consideration of the foregoing that the following hypotheses were made.

It was expected that integrated review would bring about generally higher criterion test performances at the post-acquisition and at the delayed-retention testing times. This higher performance for integrated-review students was not expected at the post-review testing time, where those students had experienced a 17-day delay prior to the post-review testing, while students in the delayed-review conditions would have just completed the review portion of the program.

The hypotheses concerning gain or difference scores would be influenced by the same factors; therefore it was expected that in most instances higher gain scores would be obtained by students having integrated review. It was however, expected that students having delayed review would obtain higher gain scores between post-acquisition and post-review.

Ability.

The literature reviewed for Experiment I indicated that interactions should have been expected between experimental treatments and students' level of mental ability. None of this literature led us to expect any such interactions with the experimental treatments in the Experiment II set.

Therefore, the simple hypothesis was made that higher ability students would have higher criterion test and gain scores, and no interactions of treatments with ability were predicted.

Time to Complete Instructional Materials.

For obvious reasons, the small-step versions of programs were hypothesized to take more time for completion than the large-step versions of programs. (However, the condensed materials for the large-step programs, because of their increased difficulty, could be expected to take longer than their reduced size might indicate.)

It was hypothesized that the delayed-review conditions would take more total work time than would integrated-review conditions. This prediction was made because the delayed-review students naturally did not have as much practice at the time of initial instruction as did the integrated-review students, and this was expected to make the delayed review materials somewhat more difficult and therefore somewhat more time consuming.

Results

Analysis of the Data.

The matching process used in assigning students to experimental treatments assured that initially there were the same number of students of each ability level in each experimental condition. However, large within treatment-group variability ensued from differences in performance between higher and lower ability students. This, with the fact that by the time of delayed-retention testing, there were no longer equal or proportional numbers of Ss in experimental conditions, led to the requirement of a multiple-classification analysis of variance correcting for disproportionate N's, as was used in Experiment I. It was decided to also use three levels of ability, in place of some other method of using ability as a control factor, so that some comparisons involving ability would be possible between this experiment and Experiment I. Data were pooled for each ability level across each experiment (all three sixth-grade classes) for each experimental program. Every analysis made included all experimental Ss and accounted for both level of ability and experimental treatments.

Results Overall.

The graphs in Figures 4, 5, and 6 present the results in percent scores for each program, all conditions, at each of the testing times.

Analyses of variance for gain scores for each program from pre-acquisition to delayed-retention show the predicted effect due to students' ability level. For the Decimals program three high ability students were eliminated from all analyses, because their pre-acquisition scores were higher than 95 percent. Nevertheless, pre-acquisition test scores were sufficiently high for a large proportion of the high ability level that a reversal in the amount gained occurred, between high and medium ability levels. High ability students had mean gains of 29.62 percent, medium ability students had mean gains of 36.88 percent, and low ability students had mean gains of 18.74 percent ($p < .05$). This analysis also indicated a significant effect for size of step ($p < .05$); small steps produced greater mean gains (33.18%) than large steps (23.65%). Table 2 gives the results of this analysis of variance.

Table 2

Analysis of Variance indicating effects measured for the Decimals program, from pre-acquisition to delayed-retention.

Source	df	MS	F	p
Review	1	122.10	1.82	<.20
Step	1	272.46	4.05	<.05*
Ability	2	333.43	4.96	<.05*
R x S	1	137.28	2.04	<.20
R x A	2	13.09		
S x A	2	77.73	1.16	
RxSxA	2	39.86		
ERROR	59	67.20		

*Indicates an effect accepted as reliable ($p < .05$)

Figure 4

Decimals Program: Results Overall

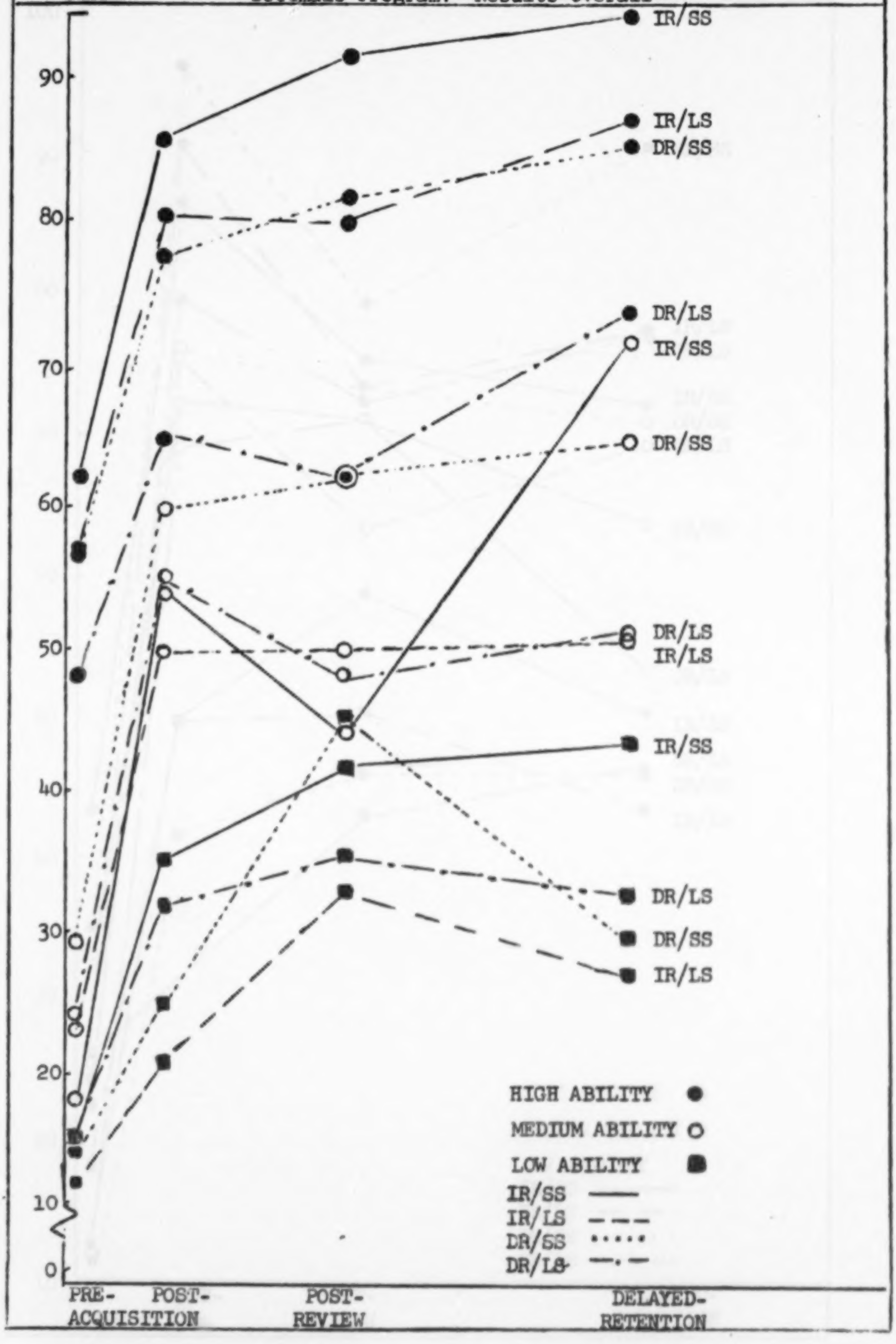


Figure 5
Squaring Program: Results Overall

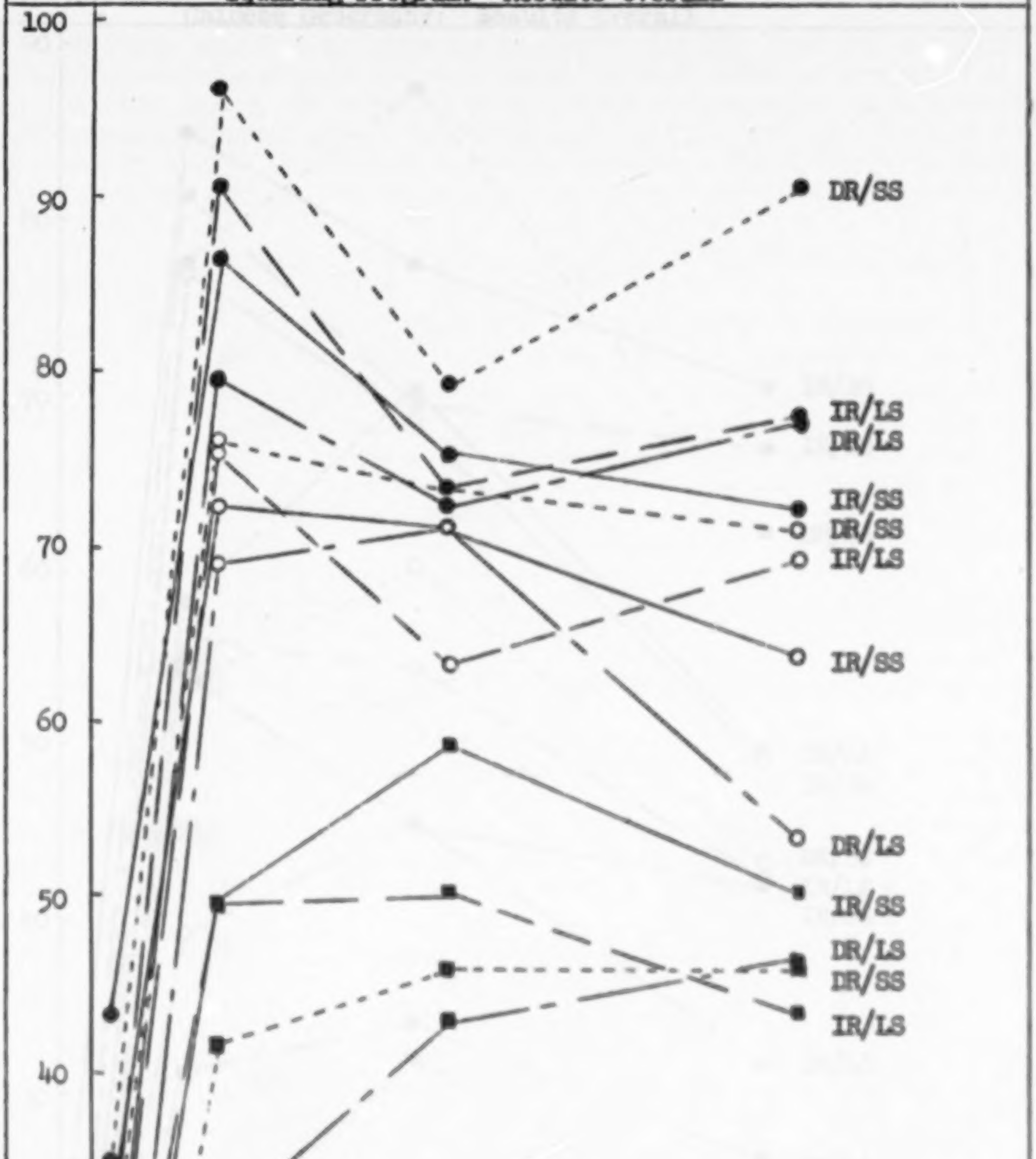
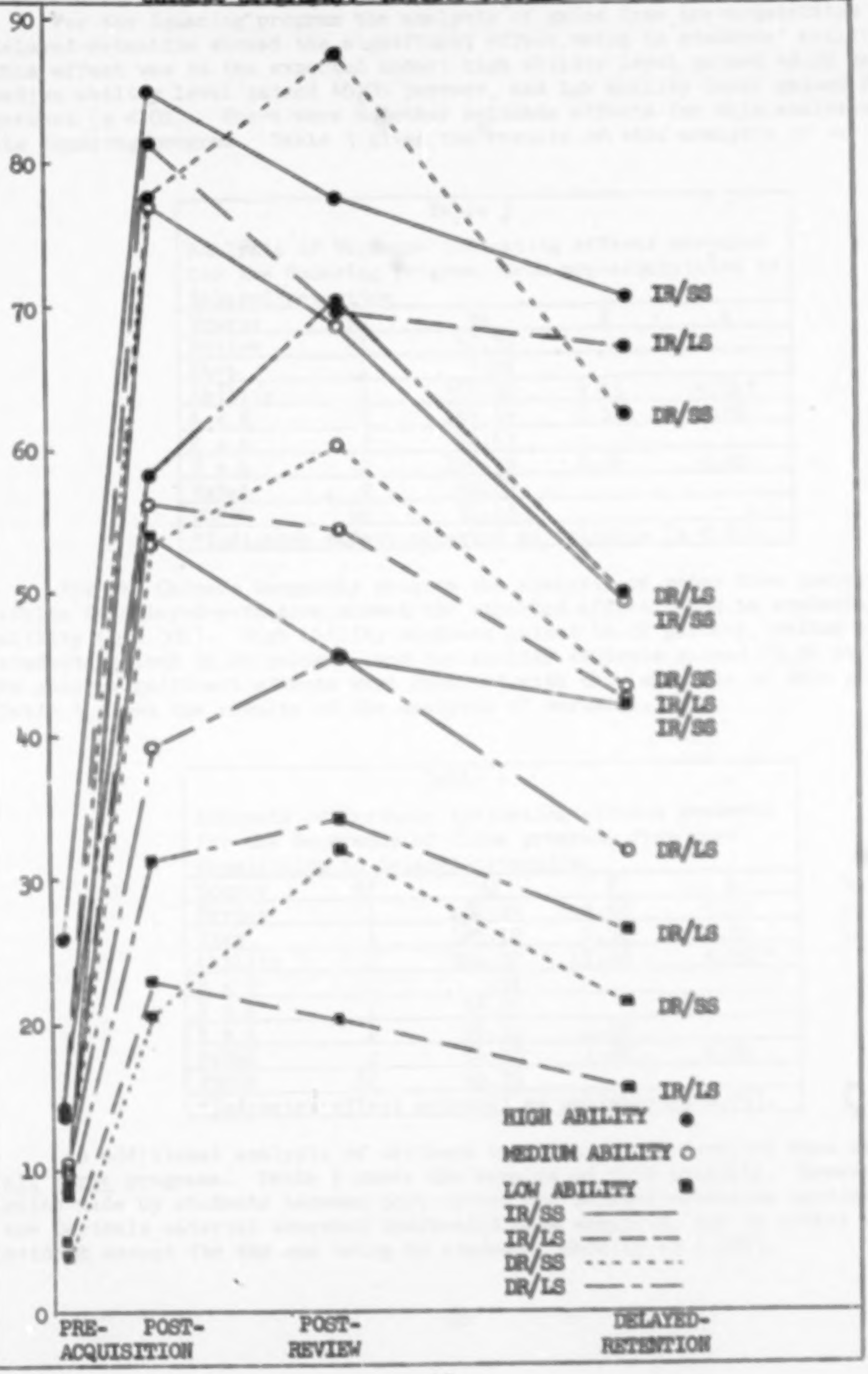


Figure 6

Chinese Geography: Results Overall



For the Squaring program the analysis of gains from pre-acquisition to delayed-retention showed the significant effect owing to students' ability level. This effect was in the expected order; high ability level gained 48.22 percent, medium ability level gained 40.20 percent, and low ability level gained 29.16 percent (p <.01). There were no other reliable effects for this analysis of the Squaring program. Table 3 gives the results of this analysis of variance.

Table 3

Analysis of Variance indicating effects measured for the Squaring program, from pre-acquisition to delayed-retention

Source	df	MS	F	p
Review	1	47.76		
Step	1	7.21		
Ability	2	366.32	5.15	<.01*
R x S	1	154.44	2.17	<.20
R x A	2	64.63		
S x A	2	146.38	2.06	<.20
RxSxA	2	49.86		
ERROR	62	71.18		

*Indicates effect accepted as reliable (p <.05).

For the Chinese Geography program the analysis of gains from pre-acquisition to delayed-retention showed the expected effect owing to students' ability (p <.001). High ability students gained 44.02 percent, medium ability students gained 32.20 percent, and low ability students gained 20.54 percent. No other significant effects were observed with this analysis of this program. Table 4 gives the results of the analysis of variance.

Table 4

Analysis of Variance indicating effects measured for the Geography of China program, from pre-acquisition to delayed-retention.

Source	df	MS	F	p
Review	1	182.99	3.68	<.10
Step	1	125.19	2.52	<.20
Ability	2	551.32	11.08	<.001*
R x S	1	.01		
R x A	2	16.25		
S x A	2	52.12	1.05	
RxSxA	2	84.28	1.69	<.20
ERROR	62	49.76		

*Indicates effect accepted as reliable (p <.05).

An additional analysis of variance was done, which involved mean data from all three programs. Table 5 shows the results of this analysis. However, the gains made by students between post-review and delayed-retention testings over the Decimals material somewhat confounded this analysis, and no effect was evident except for the one owing to students' ability (p <.005).

Table 5

Analysis of mean effects overall three programs, delayed-retention less pre-acquisition gain.

Source	df	MS	F	p
Review	1	36.72		
Step	1	67.14	1.09	
Ability	2	366.84	5.96	<.005*
R x S	1	1.69		
R x A	2	1.94		
S x A	2	73.39	1.19	
RxSxA	2	61.50	1.00	
ERROR	61	61.49		

*Indicates effect accepted as reliable (p <.05).

Figures 7 and 8, below, offer much better evidence for the support of the experimental hypotheses that (1) less effective initial instruction cannot be "made up for" by providing the same amount of redundancy or "review" at a later date (i.e., integrated review will provide better delayed review); (2) small steps will generally give better learning and retention than large steps (although somewhat different effects with this variable should be expected, depending upon the specific instructional instrument employed); (3) integrated review will produce greater gains between pre-acquisition and post-acquisition testings, while delayed-review conditions should show more gain than integrated-review conditions between post-acquisition and post-review testings.

Figure 7 shows the Small Steps conditions to have produced higher test scores at every testing time for each of the three programs--although the differences between small and large steps were not always significant.

Figure 8 shows the Integrated Review conditions to have produced higher test scores for each of the three programs at the post-acquisition testing time. Delayed Review conditions show gain between post-acquisition and post-review, and show greater gain than Integrated Review conditions for the Squaring and Geography of China programs. For the Decimals program, however, the gain between post-acquisition and post-review testings is about equal for both Integrated- and Delayed-Review conditions.

At the delayed retention testing time the hypotheses about timing of review are strictly upheld only by the results with the Geography of China program; namely, Integrated-Review conditions show higher absolute scores and less loss than Delayed-Review conditions. For the Squaring program the Delayed-Review conditions appear to be slightly superior (about 1.39 higher) to Integrated-Review conditions.

For the Decimals program, the effect of ongoing classroom instruction in mathematics is apparent--even though none of the teachers had formally introduced decimals by the time of our delayed-retention testing of that subject-matter. This gain between post-review and delayed-retention testing times is attributed by the authors to ongoing classroom instruction (although there are alternate explanations, such as, the post-review test was easier than the delayed-

Figure 7

Small steps versus large steps, each experimental program overall.

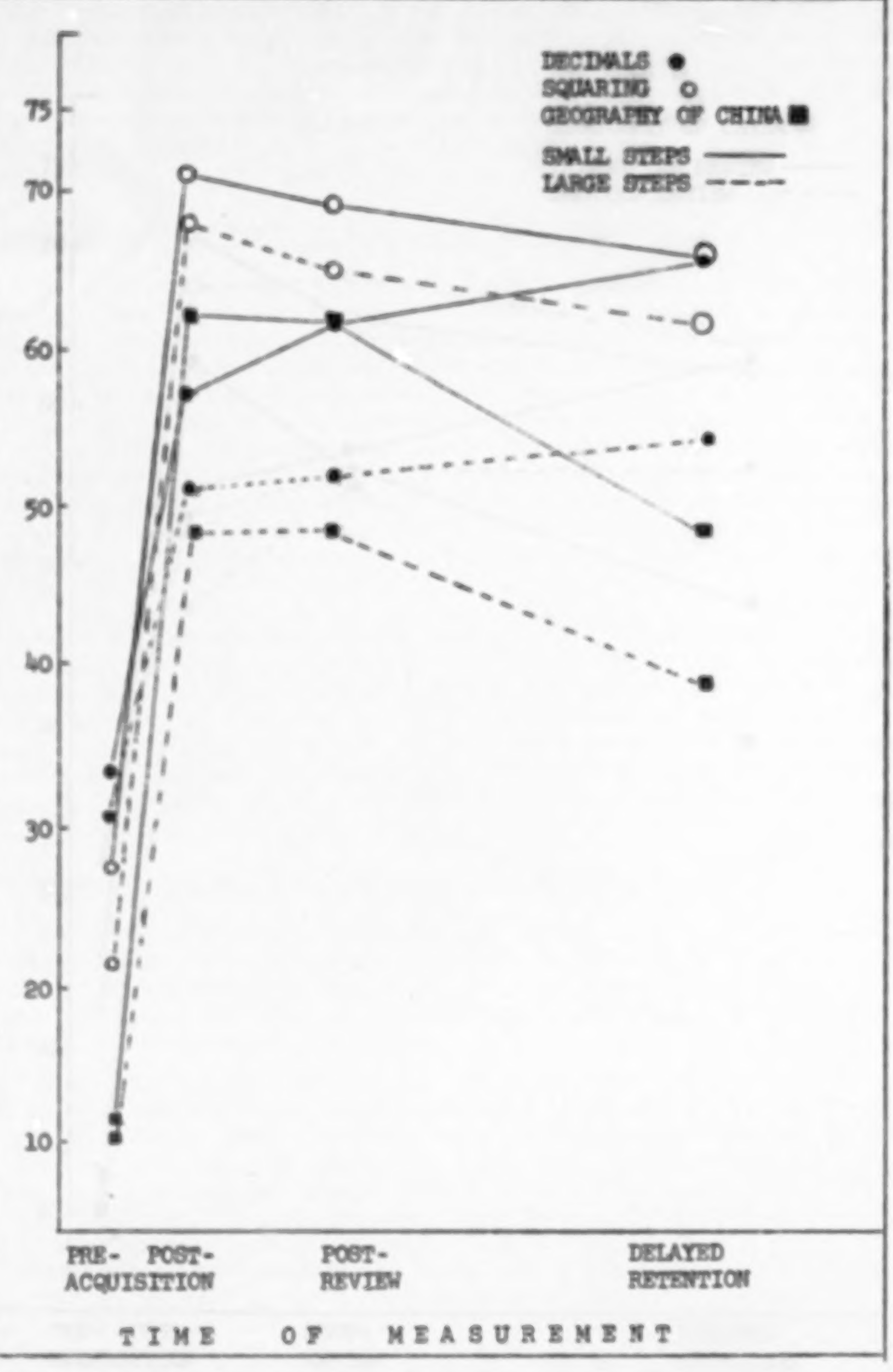
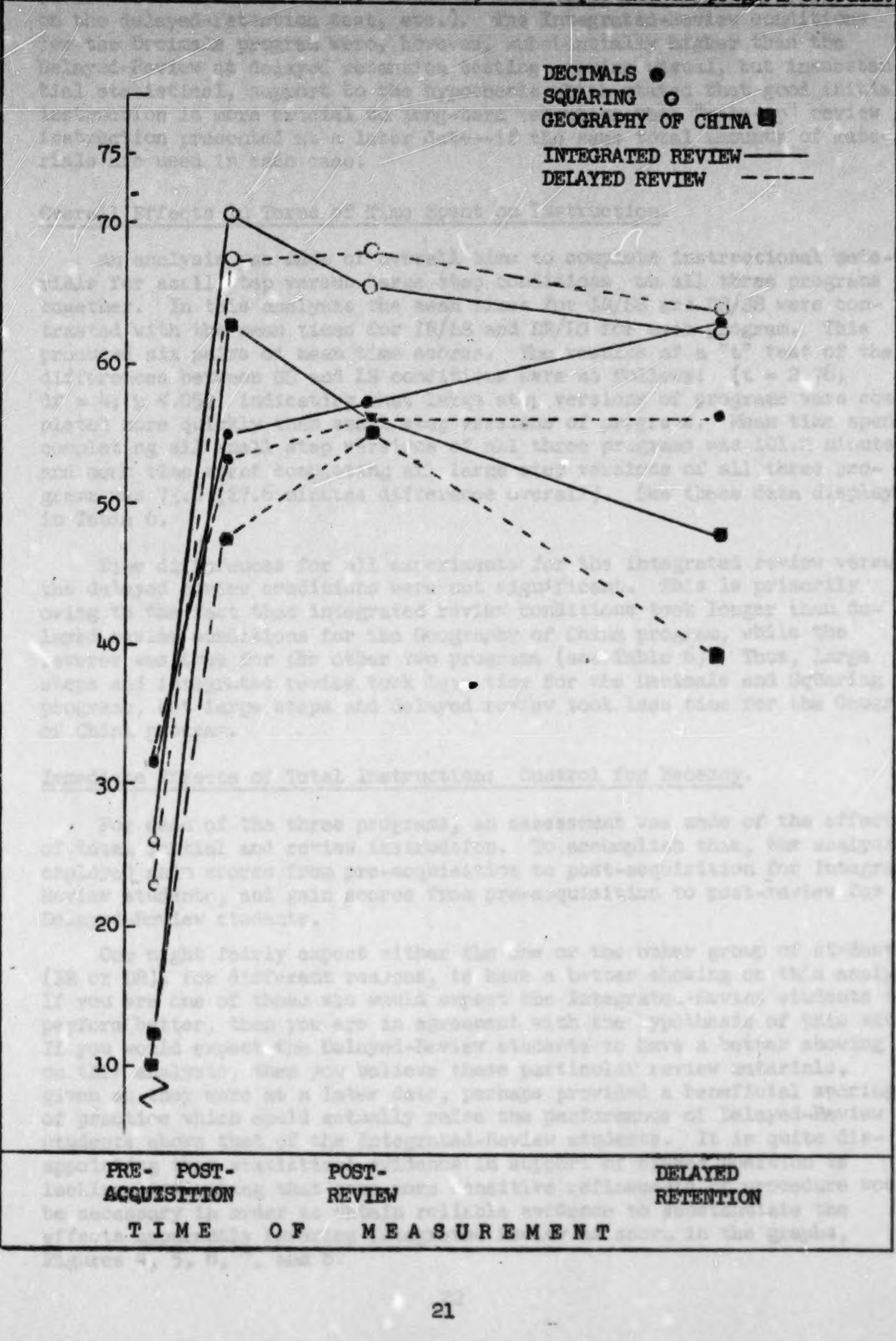


Figure 8

Integrated review versus delayed review, each experimental progrm overall.



retention test; or the students learned something from taking the post-review test which was not reflected on that test, but which was reflected on the delayed-retention test, etc.). The Integrated-Review conditions for the Decimals program were, however, substantially higher than the Delayed-Review at delayed retention testing, giving visual, but insubstantial statistical, support to the hypothesis which stated that good initial instruction is more crucial to long-term retention than "make-up" review instruction presented at a later date--if the same total amounts of materials are used in each case.

Overall Effects in Terms of Time Spent on Instruction.

An analysis was made of overall time to complete instructional materials for small step versus large step conditions, on all three programs together. In this analysis the mean times for IR/SS and DR/SS were contrasted with the mean times for IR/LS and DR/LS for each program. This produced six pairs of mean time scores. The results of a "t" test of the differences between SS and LS conditions were as follows: ($t = 2.78$; $df = 4$; $p < .05$), indicating that large step versions of programs were completed more quickly than small step versions of programs. Mean time spent completing all small step versions of all three programs was 101.2 minutes and mean time spent completing all large step versions of all three programs was 73.6 (27.6 minutes difference overall). See these data displayed in Table 6.

Time differences for all experiments for the integrated review versus the delayed review conditions were not significant. This is primarily owing to the fact that integrated review conditions took longer than delayed review conditions for the Geography of China program, while the reverse was true for the other two programs (see Table 6). Thus, large steps and integrated review took less time for the Decimals and Squaring programs, but large steps and delayed review took less time for the Geography of China program.

Immediate Effects of Total Instruction: Control for Recency.

For each of the three programs, an assessment was made of the effects of total initial and review instruction. To accomplish this, the analysis employed gain scores from pre-acquisition to post-acquisition for Integrated-Review students, and gain scores from pre-acquisition to post-review for Delayed-Review students.

One might fairly expect either the one or the other group of students (IR or DR), for different reasons, to have a better showing on this analysis. If you are one of those who would expect the Integrated-Review students to perform better, then you are in agreement with the hypothesis of this study. If you would expect the Delayed-Review students to have a better showing on this analysis, then you believe these particular review materials, given as they were at a later date, perhaps provided a beneficial spacing of practice which could actually raise the performance of Delayed-Review students above that of the Integrated-Review students. It is quite disappointing that statistical evidence in support of either position is lacking--indicating that some more sensitive refinements of procedure would be necessary in order to obtain reliable evidence to substantiate the effects apparently favoring Integrated Review as shown in the graphs, Figures 4, 5, 6, 7, and 8.

Table 6

Number of frames in each program and condition; mean time in minutes to completion, and mean gain from pre-acquisition to delayed-retention. Test scores are per cents correct.

Program and Step size		Integrated Review					Delayed Review				
		Number of Frames:		Mean Time for Completion of All Instruction	Mean Time for Completion of and Gain Scores on Delayed-Retention Test		Number of Frames:		Mean Time for Completion of All Instruction	Mean Time for Completion of and Gain Scores on Delayed-Retention Test	
		Initial	Review		Time	Gain	Initial	Review		Time	Gain
Decimals Program	SS	209	0	97.94	27.16	39.75	76	133	104.33	24.11	26.60
	LS	156	0	91.45	22.79	23.46	23	133	93.72	25.16	23.84
Squaring Program	SS	54	0	24.18	11.14	32.84	32	22	24.86	10.39	44.00
	LS	42	0	22.08	12.63	41.57	20	22	23.53	12.87	38.38
Chinese Geography Program	SS	196	0	134.99	31.79	39.41	91	105	120.78	25.79	31.55
	LS	143	0	110.91	29.41	32.90	38	105	99.80	24.11	25.15

For the Decimals program, a significant effect owing to step size ($p < .05$) is present in this control for recency analysis. The direction of the effect is such that small steps produced gain of 28.47 (per cent scores) and large steps produced gain of only 19.26. There were no other discriminable effects, which is in itself worthy of comment. The effect which one would ordinarily expect, owing to level of ability, is undoubtedly missing here because of the ceiling effect which occurs with this subject-matter in the high ability level--that is, many pre-acquisition test scores for high ability Ss were greater than 80 per cent. See Table 7 for results of this analysis of variance.

For the Squaring program, a significant effect owing only to level of ability ($p < .025$) is present in the control for recency analysis; such that high ability students gained 51.0 (per cent scores), medium ability students gained 48.9, and low ability students gained 35.5. See Table 8 for the results of this analysis.

For the Geography of China program, significant effects owing both to step size ($p < .01$) and to ability level ($p < .001$) are present in the control for recency analysis. High ability students gained 64.0 (per cent scores), medium ability students gained 50.0, and low ability students gained 30.1. Small step conditions showed gain of 54.4, as opposed to large step conditions, which showed gain of 41.7. See Table 9 for the results of this analysis.

Table 7

Analysis of Variance, control for recency, for the Decimals program: Gains from pre-acquisition to post-acquisition for Integrated-Review students and gains from pre-acquisition to post-review for Delayed-Review students.

Source	df	MS	F	p
Review	1	.58		
Step	1	254.46	4.91	<.05*
Ability	2	83.96	1.62	
R x S	1	17.86		
R x A	2	66.93	1.29	
S x A	2	13.03		
RxSxA	2	7.12		
ERROR	59	51.83		

*Indicates effect accepted as reliable ($p < .05$)

Table 8

Analysis of Variance, control for recency, Squaring program: Gains from pre-acquisition to post-acquisition for Integrated-Review students, and gains from pre-acquisition to post-review for Delayed-Review students.

Source	df	MS	F	p
Review	1	.10		
Step	1	119.46	1.79	<.20
Ability	2	282.52	4.23	<.025*
R x S	1	39.30		
R x A	2	79.48	1.19	
S x A	2	142.76	2.14	<.20
RxSxA	2	39.35		
ERROR	62	66.82		

Table 9

Analysis of Variance, control for recency, Geography of China program: Gains from pre-acquisition to post-acquisition for Integrated-Review students, and gains from pre-acquisition to post-review for Delayed-Review students.

Source	df	MS	F	p
Review	1	97.44	1.60	
Step	1	481.56	7.92	<.01*
Ability	2	1159.38	19.07	<.001*
R x S	1	.48		
R x A	2	69.05	1.14	
S x A	2	44.96		
RxSxA	2	144.57	2.38	<.20
ERROR	62	60.79		

---

*Indicates an effect accepted as reliable (p <.05).

Additional Analyses.

A number of analyses were made which were peripheral to the central purpose of assessing effects on long-term retention of knowledge. These analyses are given in Appendix A. Tables showing critical differences[1] between means are also given there, wherever an interaction effect reached or exceeded the .05 level of probability.

Some of these analyses point out really interesting things about changes that took place over different periods of time--different periods of instruction or review versus periods of no instruction or review, etc.--, and differences in effects between the three programs. The most obvious conclusion is that effects for the three programs are quite different--even though the same experimental design, subjects, and "variables" were employed for each program. This provides eloquent evidence that the program itself is a greater source of variability than any of the so-called "variables" which have been often employed (as they were here) in research with programed instruction.

Discussion and Conclusions

Most of the overall results of this study are fairly easy to interpret. Statistical evidence supports the hypothesis that higher ability students (a) usually know more to begin with than low ability students; and (b) generally learn more (gain more) than low ability students, although there were exceptions to this generalization, mainly due to a ceiling effect observed with the Decimals materials. Results also indicated that size of step was a crucial factor in retention with only the Decimals program. However, size of step appeared to be crucial for Geography of China in terms of immediate effects of instruction.

If measures of gains and obtained scores are examined for periods of time inside of the pre-acquisition to delayed-retention span (changes between post-review and delayed-retention, or changes between post-acquisition and post-review, for example), other experimental variations assume importance.

The hypotheses about recency effects were substantiated, at least by graphic evidence (Figures 7 and 8), but other than the recency effect, no reliable evidence was obtained for the Integrated- versus Delayed-Review conditions.

There were many more reliable findings with the Geography of China materials than with either of the other two programs--although the levels of initial learning and retention were generally lower for this program than for either of the other two. The Decimals program was next in order for number of significant effects, and learning and retention of that material was next above that of Chinese Geography. The Squaring program had fewest significant effects, and learning and retention of that material was highest of all. In connection with these particular findings a number of possibilities might be discussed:

---

[1] Lindquist, E. F. Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin, 1953, p. 93.

(1) One might decide that the program which had been most extensively tested and revised was the one which would produce highest scores, greatest gain, and greatest retention. This argument needs to be further investigated, since it is obviously unfair to the Geography of China program, which is quite different in subject-matter type and content design from the other two programs In connection with this, the reader should also know that the Geography of China program was intended by its authors for older students, and if this is to be taken as indicative of its level of difficulty, one might wish to argue that it was the superior program, since these sixth-grade students performed nearly as well with it as with the other "more nearly sixth-grade level" programs.

(2) One might decide that the most crucial part of the learning for mathematical subject-matters such as the Decimals and Squaring programs contained, occurs at initial acquisition time--because the learning involves only a very few closely related concepts, based one on the next; while for the Chinese Geography materials, additional material may be learned at any time, since the knowledge involves fairly discrete and fairly unrelated facts and concepts, not necessarily based one on the next. Further research to investigate these types of subject-matter relationships would clarify many of the now-existing apparent conflicts among other research findings.

(3) One might also argue that the best learning occurred with the program for which the behavioral objectives were both fewest and easiest to set, namely squaring two-digit numbers ending in five. The least learning then can be argued to have occurred with the Geography of China program, for the reason that behavioral objectives for that subject-matter were both most numerous and most difficult to set (or at least to justify having set). This is certainly a logical notion, if one accepts the implication that the program is not quite as good when many difficult decisions have to be made about what the behavioral objectives should be. Further research could clarify this point.

(4) One might ignore the overall results from pre-acquisition to delayed-retention testing, and look at the internal or more short-term effects with the Geography of China program. By far the largest number of significant effects occurred with this program, and the argument could be given that this material was the most difficult, and that therefore the variables employed in the study--which are variables very often discussed in connection with programed instruction--made more difference with this material than with Decimals or Squaring.

(5) One can suggest that the Chinese Geography program was the most difficult, had the highest ceiling, and therefore was logically going to produce slightly lower criterion test scores.

(6) The Squaring program was undoubtedly the least dense of the three, the Decimals program was the next least dense, and the Geography of China program was the most dense--in terms of both the spacing of the practice provided by content steps, and in terms of the number of discrete facts which had to be learned, most of which did not prompt recall of other facts. The argument that program density is a determiner of final effects--regardless of the amount of time a student might spend on a dense or less dense program--, could therefore be evaluated by further research. This argument is related to the size-of-step as spacing-of-practice discussion put forth in the rationale section of the present study. Our results definitely suggest that spaced practice in the form of less dense subject-matter has an effect both on how much the student learn and gains, and on what other variations in the instructional situation, materials content, and materials format might be expected to produce reliable effects. Whether the spacing of practice which is beneficial to the student is provided by reduced content density of small steps, as happens in the usual linear program, or whether the beneficial spacing of practice is that which is provided simply by small context or program format steps is a question not fully evaluated in the present research, and provides a direction for further investigation.

In summary, then, the programs all showed different results. The level of ability of the students produced the most consistent main effect. The size of step variable was definitely important for the Decimals materials, was probably important for the Chinese Geography materials, and did not seem to have any pronounced effect for the Squaring materials. Overall, the time of review variable (Integrated- versus Delayed-) showed no lasting effects for any of the three programs. The Small Steps instructional materials took significantly longer to complete than the Large Steps materials, but no great variations in time of completion could be related to Integrated- versus Delayed-Review.

In contrast to the results of Experiment I and to the literature reviewed for Experiment I in the present project, the most striking finding was that not only did students of different abilities gain more or less but, over periods of no treatment (no instruction no review), students of different ability levels lost differential amounts.

References

Angell, D. & Campeau, P. L. Squaring two-digit numbers ending in five. An unpublished program developed for research purposes and based upon an earlier program on the same topic, written by J. L. Evans of Teaching Machines Incorporated, Albuquerque, New Mexico, 1962.

Angell, D. & Lumsdaine, A. A. Retention of material presented by autoinstructional programs which vanish and which do not vanish verbal cues. San Mateo, Calif.: American Institute for Research, 1962. AIR-C14-8/62-TR.

Brown, D. R., Naylor, J. C., & Briggs, G. E. The retention of discrete and continuous tasks as a function of interim practice with modified task requirements. USAF AMRL - TDR, 1963, No. 63-35.

Coulson, J. E. & Silberman, H. F. Effects of 3 variables in a teaching machine. J. educ. Psychol., 1960, 51, 135-43.

Evans, J. L., Glaser, R., & Homme, L. E. A preliminary investigation of variation in the properties of verbal learning sequences of the "Teaching Machine" type. In A. A. Lumsdaine and R. Glaser, Teaching machines and programmed learning. National Education Association of U.S., Wash., D. C., 1960, 446-451.

Ferster, L. B. The role of review material in continuous programing with teaching machines. Indianapolis, Indiana Univer. Med. School, 1960. In W. Schramm, Research on programmed instruction, Inst. for Communication Res., Stanford Univer., 1962.

Gagné, R. M. The Conditions of Learning. New York: Holt, Rinehart and Winston, Inc., 1964.

Glaser, R. & Reynolds, J. H. Investigation of learning variables in programmed instruction; Part I: Repetition and spaced review in programmed instruction. AF 33(616)-7175 and OE 2-10-057, Univer. Pittsburgh, Pa., October, 1962.

Hamilton, Nancy R., Willis, Mary B., Bivens, L. W., & Doolin, D. J. Learning about Communism. Palo Alto, Calif.: American Institutes for Research, 1963.

Holland, J. G. & Porter, D. The influence of repetition of incorrectly answered items in a teaching-machine program. J. exp. Analysis Behav., 1961, 4, (4), 305-307. In W. Schramm, Research on programmed instruction. Inst. for Communication Res., Stanford Univer., 1962.

Holland, J. G. & Skinner, B. F. The Analysis of Behavior. New York: McGraw-Hill, 1961.

Lumsdaine, A. A., Sulzer, R. L., & Kopstein, F. F. The effect of animation cues and repetition of examples on learning from an instructional film. In W. Schramm, Research on programmed instruction. Inst. for Communication Res., Stanford Univer., 1962.

Naylor, J. C., Briggs, G. E., & Reed, W. G. The effects of task organization, training time, and retention interval on the retention of skill. USAF AMRL - TDR, 1962, No. 62-107.

Schramm, W. Research on programmed instruction. An annotated bibliography. Inst. for Communication Res., Stanford Univer., 1962.

Smith, M. D. Decimals and per cents. Belmont, Calif.: Allyn and Bacon, 1964.

Smith, W. & Moore, J. W. Size-of-step and achievement in programmed spelling. Lewisburg, Pa.: Bucknell Univer., 1961.

Soens, T., Harr, E., Mavrinac, H., Rifugiato, F., Kline, Jr., H. V. B., & Dreibelbis, P. China: A programmed unit in geography. New York: Holt, Rinehart and Winston, 1963.

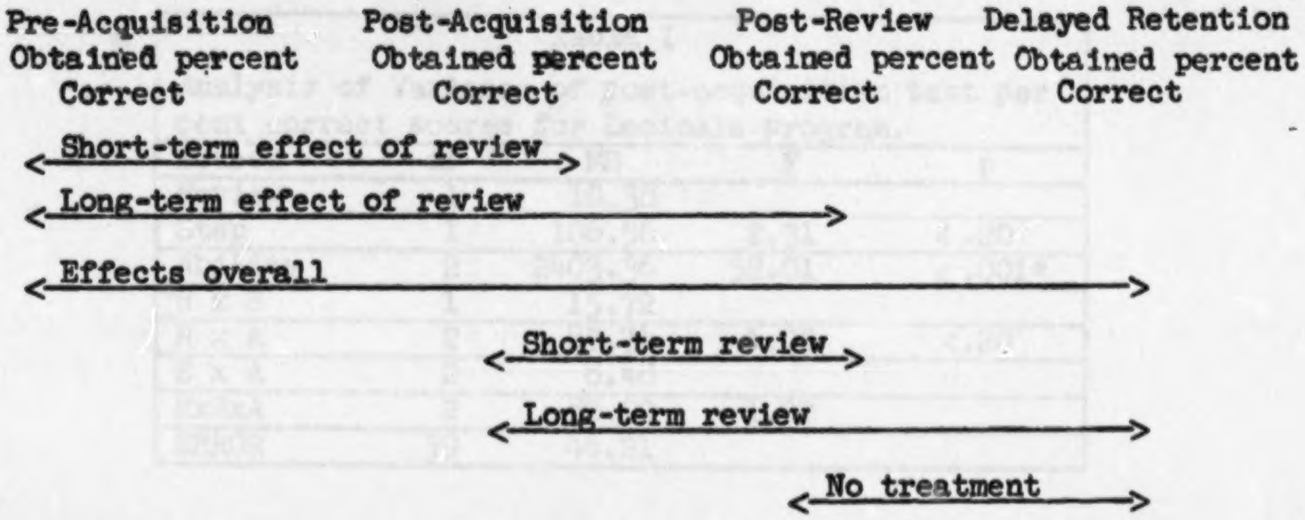
Underwood, B. J. Ten years of massed practice on distributed practice. Psychol. Rev., 1961, 68, (4), 229-247.

Underwood, B. J. Forgetting, Scientific American, 1964, 210, (3), 91-99.

Appendix A

Additional Analyses

Appendix A


Pre-Acquisition Obtained percent Correct
Post-Acquisition Obtained percent Correct
Post-Review Obtained percent Correct
Delayed Retention Obtained percent Correct
Short-term effect of review
Long-term effect of review
Effects overall
Short-term review
Long-term review
No treatment

Analyses were made for each program at the above times. Arrows indicate differences analyses.

Table I

Analysis of Variance of post-acquisition test per cent correct scores for Decimals program.

Source	df	MS	F	p
Review	1	10.38		
Step	1	106.56	2.31	< .20
Ability	2	2403.36	52.01	< .001*
R x S	1	15.72		
R x A	2	87.74	1.90	<.20
S x A	2	8.48		
RxSxA	2	82.90	1.22	
ERROR	59	46.21		

Table II

Analysis of Variance of post-acquisition test percent correct scores for Squaring program.

Source	df	MS	F	p
Review	1	37.81		
Step	1	22.20		
Ability	2	1880.72	31.42	<.001*
R x S	1	78.95	1.32	
R x A	2	16.64		
S x A	2	10.51		
RxSxA	2	27.20		
ERROR	62	59.85		

Table III

Analysis of Variance of post-acquisition test per cent correct scores for Chinese Geography program.

Source	df	MS	F	p
Review	1	747.02	14.87	<.001*
Step	1	529.81	10.54	<.01*
Ability	2	1842.18	36.67	<.001*
R x S	1	83.16	1.66	<.20
R x A	2	15.71		
S x A	2	13.64		
RxSxA	2	203.76	4.06	<.05*
ERROR	62	50.24		

---

*Indicates an effect accepted as significant ($p < .05$).

Mean percent correct scores for significant effect ($p < .001$) due to ability level on the post-acquisition analysis of the Decimals program:

HA	77.87
MA	55.38
LA	28.90

- -

Mean percent correct scores for significant effect ($p < .001$) due to ability level on the Post-acquisition analysis of the Squaring program:

HA	88.30
MA	73.35
LA	45.57

- -

Mean percent correct scores for significant main effects on the post-acquisition analyses of the Geography of China program:

Review ($p < .001$)	IR	62.84
	DR	47.06
Step ($p < .01$)	SS	61.60
	LS	48.31
Ability ($p < .001$)	HA	75.56
	MA	56.56
	LA	32.73

Table IV

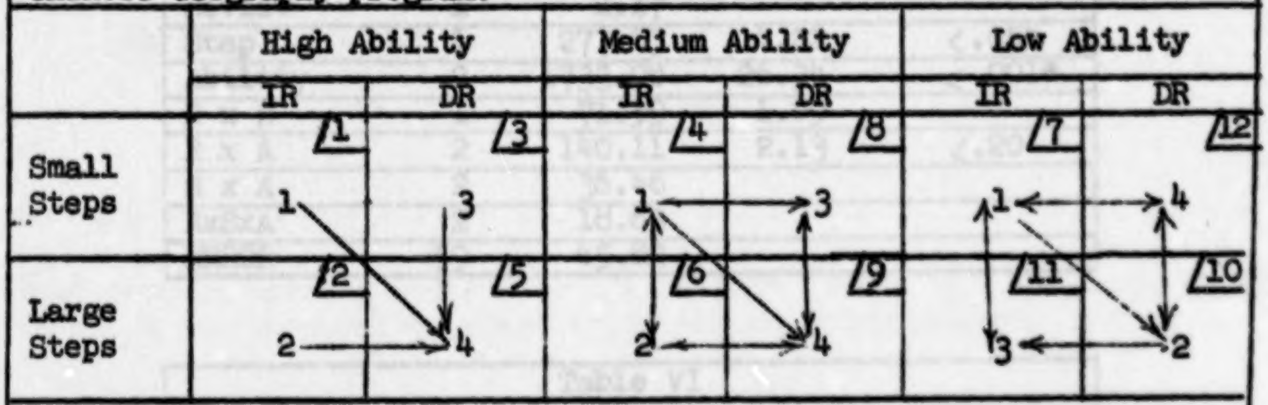
Means and Critical Differences for Step x Review x Ability Interaction ($p < .05$) at Post-Acquisition on the Geography of China Program. Critical Difference for $p = .05$ on this Analysis is 8.18.

Ability				High Ability				Medium Ability				Low Ability			
	Review			IR		DR		IR		DR		IR		DR	
		Step		SS	LS	SS	LS	SS	LS	SS	LS	SS	LS	SS	LS
			means	85.25	81.28	77.50	58.21	76.71	56.33	53.93	39.28	54.39	23.09	21.79	31.66
High	IR	SS	85.25		3.97	7.75	27.04*	8.54*	28.92*	31.32*	45.97*	30.86*	62.16*	63.46*	53.59*
		LS	81.28			3.78	23.07*	4.57	24.95*	27.35*	42.00*	26.89*	58.19*	59.49*	49.62
	DR	SS	77.50				19.29*	0.79	21.17*	23.57*	38.22*	23.11*	54.41*	55.71*	45.84*
		LS	58.21					-18.50*	1.88	4.28	18.93*	3.82	35.12*	36.42*	26.55
Medium	IR	SS	76.71						20.38*	22.78*	37.43*	22.32*	53.62*	54.92*	45.05*
		LS	56.33							2.40	17.05*	1.94	33.24*	34.54*	24.67*
	DR	SS	53.93								14.65*	-0.46	30.84*	32.14*	22.27*
		LS	39.28									-15.11*	16.19*	17.49*	7.62
Low	IR	SS	54.39										31.30*	32.60*	22.73*
		LS	23.09											1.30	-8.54*
	DR	SS	21.79												-9.87*
		LS	31.66												

* $p \leq .05$

Figure I

Rank Orders: R x S x A Interaction, post-acquisition means on Chinese Geography program.


High Ability
Medium Ability
Low Ability
IR
DR
IR
DR
IR
DR
Small Steps
Large Steps
1
3
4
8
7
12
1
3
1
3
1
4
2
5
6
9
11
10
2
4
2
4
3
2

Arrows indicate significant differences between treatments within each ability level. (←→) The numbers in the large boxes reflect the rank order of the treatments within each ability level. The number enclosed in the upper right corner of each box reflects the position of that cell over the entire interaction. A rank of 1 indicates the best treatment.

Table V

Analysis of Variance of post-review test per cent correct scores for Decimals program.

Source	df	MS	F	p
Review	1	2.17		
Step	1	277.63	4.22	$<.05$*
Ability	2	1733.98	26.34	$<.001$*
R x S	1	71.59	1.09	
R x A	2	140.11	2.13	$<.20$
S x A	2	36.46		
RxSxA	2	18.63		
ERROR	59	65.82		

Table VI

Analysis of Variance of post-review test per cent correct scores for Squaring program.

Source	df	MS	F	p
Review	1	20.91		
Step	1	48.00	2.23	$<.20$
Ability	2	432.25	20.11	$<.001$*
R x S	1	14.59		
R x A	2	5.52		
S x A	2	2.18		
RxSxA	2	19.06		
ERROR	62	21.49		

Table VII

Analysis of Variance of post-review test per cent correct scores for Chinese Geography program.

Source	df	MS	F	p
Review	1	4.25		
Step	1	487.69	7.13	$<.01$*
Ability	2	1876.17	27.42	$<.001$*
R x S	1	23.18		
R x A	2	50.58		
S x A	2	2.36		
RxSxA	2	89.83	1.31	
ERROR	62	68.43		

---

*Indicates effect accepted as reliable ($p<.05$).

Mean percent correct scores for significant main effects on the post-review analyses of the Decimal program:

Step ($p < .05$)	SS	61.59
	LS	51.97
Ability ($p < .001$)	HA	79.72
	MA	51.54
	LA	39.08

- -

Mean percent correct scores for significant effect ($p < .001$) due to ability level on the post-review analysis of the Squaring program:

HA	75.42
MA	70.01
LA	55.33

- -

Mean percent correct scores for significant main effects on the post-review analyses of the Geography of China program:

Step ($p < .01$)	SS	61.99
	LS	49.24
Ability ($p < .001$)	HA	76.40
	MA	57.27
	LA	33.18

Table VIII

Analysis of Variance of delayed-retention test per cent correct scores for Decimals program.

Source	df	MS	F	p
Review	1	114.21	1.50	
Step	1	372.30	4.48	<.05*
Ability	2	2733.20	32.87	<.001*
R x S	1	43.20		
R x A	2	19.24		
S x A	2	33.28		
RxSxA	2	34.95		
ERROR	59	83.14		

Table IX

Analysis of Variance of delayed-retention test per cent correct scores for Squaring program.

Source	df	MS	F	p
Review	1	5.80		
Step	1	54.70	1.10	
Ability	2	1073.21	21.68	<.001*
R x S	1	102.14	2.06	<.20
R x A	2	45.54		
S x A	2	2.43		
RxSxA	2	62.82	1.27	
ERROR	62	49.50		

Table X

Analysis of Variance of delayed-retention test per cent correct scores for Chinese Geography program.

Source	df	MS	F	p
Review	1	227.04	4.79	< .05*
Step	1	257.82	5.44	< .025*
Ability	2	1293.58	27.28	<.001*
R x S	1	26.64		
R x A	2	14.13		
S x A	2	2.85		
RxSxA	2	119.92	2.53	<.10
ERROR	62	47.42		

---

*Indicates effect accepted as reliable ($p < .05$).

Mean percent correct scores for significant main effects on the delayed-retention analysis of the Decimals program:

Step ($p < .05$)	SS	65.58
	LS	54.44
Ability ($p < .001$)	HA	86.16
	MA	59.99
	LA	33.88

- -

Mean percent correct scores for significant effect ($p < .001$) due to ability level on the delayed-retention analysis of the Squaring program:

HA	79.43
MA	64.47
LA	46.71

- -

Mean percent correct scores for significant main effects on the delayed-retention analysis of the Geography of China program:

Review ($p < .05$)	IR	47.97
	DR	39.27
Step ($p < .025$)	SS	48.26
	LS	38.99
Ability ($p < .001$)	HA	62.41
	MA	41.88
	LA	26.57

Table XI

Analysis of Variance, pre-acquisition to post-acquisition gains for Decimals program.

Source	df	MS	F	p
Review	1	14.52		
Step	1	25.23		
Ability	2	267.86	6.09	<.005*
R x S	1	44.93	1.02	
R x A	2	8.09		
S x A	2	3.04		
RxSxA	2	30.39		
ERROR	59	44.01		

Table XII

Analysis of Variance, pre-acquisition to post-acquisition gains for Squaring program.

Source	df	MS	F	p
Review	1	0.12		
Step	1	35.50		
Ability	2	821.80	11.55	<.005*
R x S	1	126.16	1.77	<.20
R x A	2	92.34	1.30	
S x A	2	114.32	1.61	<.20
RxSxA	2	63.18		
ERROR	62	71.14		

Table XIII

Analysis of Variance, pre-aquisition to post-acquisition gains for Chinese Geography program.

Source	df	MS	F	p
Review	1	447.98	9.44	<.001*
Step	1	516.40	10.88	<.001*
Ability	2	878.77	18.52	<.001*
R x S	1	0.00		
R x A	2	58.15	1.22	
S x A	2	37.71		
RxSxA	2	155.68	3.28	<.05*
ERROR	62	47.46		

---

*Indicates effect accepted as reliable (p < .05).

Mean gain for significant effect ($p < .005$) due to ability level on the pre-acquisition to post-acquisition differences analysis of the Decimals program:

Gain:		
	HA	21.32*
	MA	31.26*
	LA	15.03

*Notice the reversal in expected order, owing to high pre-acquisition scores made by High Ability Ss.

Mean gain for significant effects ($p < .005$) due to ability level on the pre-acquisition to post-acquisition differences analysis of the Squaring program:

Gain:		
	HA	56.89
	MA	48.85
	LA	29.04

Mean gain scores for significant main effects on the pre-acquisition to post-acquisition differences analysis of the Geography of China program:

	Review ($p < .001$)	IR	50.89
		DR	38.67
Gain:	Step ($p < .001$)	SS	51.44
		LS	38.22
	Ability ($p < .001$)	HA	58.44
		MA	46.88
		LA	29.02

Table XIV

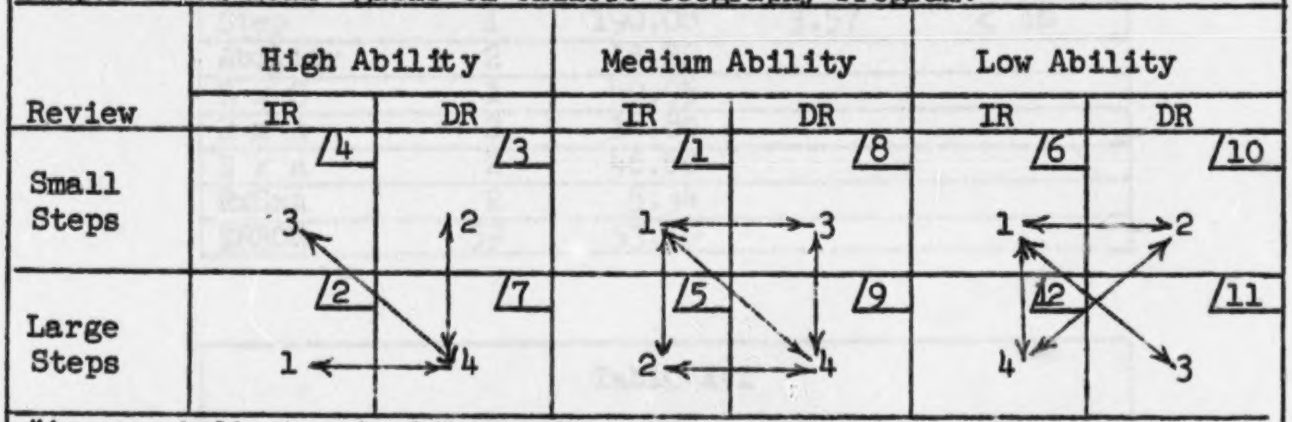
Means and Critical Differences for Step x Review x Ability Interaction ($p < .05$), for Post-Acquisition less Pre-Acquisition Mean Gains on the Geography of China Program. Critical Difference for $p = .05$ is 7.96.

Ability			High Ability				Medium Ability				Low Ability			
	Review		IR		DR		IR		DR		IR		DR	
		Step	SS	LS	SS	LS	SS	LS	SS	LS	SS	LS	SS	LS
		Means	59.25	66.93	63.83	43.76	68.00	46.62	43.68	29.21	45.05	19.48	28.22	23.33
High	IR	SS 59.25		-7.68	-4.58	15.49*	-8.75*	12.63*	15.57*	30.04*	14.20*	39.77*	31.03*	35.92*
		LS 66.93			3.10	23.17*	-1.07	20.31*	23.25*	37.72*	21.88*	47.45*	38.71*	43.60*
	DR	SS 63.83				20.07*	-4.17	17.21*	20.15*	34.62*	16.78*	44.35*	35.61*	40.50*
		LS 43.76					-24.24*	-2.86	.08	14.55*	-1.29	24.28*	15.54*	20.43*
Medium	IR	SS 68.00						21.38*	24.32*	38.79*	22.95*	48.52*	39.78*	44.67*
		LS 46.62							2.94	17.41*	1.57	27.14*	18.40*	23.29*
	DR	SS 43.68								14.47*	-1.37	24.20*	15.46*	20.35*
		LS 29.21									-15.84*	9.73*	.99	5.88
Low	IR	SS 45.05										25.57*	16.83*	21.72*
		LS 19.48											-8.74*	-3.85
	DR	SS 28.22												4.89
		LS 23.33												

*$p \leq .05$

Figure II

Rank Orders: R x S x A Interactions, post-acquisition less pre-acquisition mean gains on Chinese Geography Program.



	High Ability		Medium Ability		Low Ability	
Review	IR	DR	IR	DR	IR	DR
Small Steps	4 3	3 2	1 1	8 3	6 1	10 2
Large Steps	2 1	7 4	5 2	9 4	12 4	11 3

*Arrows indicate significant differences between treatments within each ability level. (←——→) The numbers in the large boxes reflect the rank order of the treatments within each ability level. The number enclosed in the upper right corner of each box reflects the position of that cell over the entire interaction. A rank of 1 indicates the best treatment.

Table XV

Analysis of Variance, pre-acquisition to post-review gains for Decimals program.

Source	df	MS	F	p
Review	1	11.76		
Step	1	190.08	3.57	<.10
Ability	2	15.22		
R x S	1	40.95		
R x A	2	18.80		
S x A	2	46.60		
RxSxA	2	5.34		
ERROR	59	53.19		

Table XVI

Analysis of Variance, pre-acquisition to post-review gains for Squaring program.

Source	df	MS	F	p
Review	1	92.07	1.85	<.20
Step	1	13.48		
Ability	2	67.02	1.34	
R x S	1	0.98		
R x A	2	74.98	1.50	
S x A	2	139.85	2.81	<.10
RxSxA	2	56.50	1.13	
ERROR	62	49.80		

Table XVII

Analysis of Variance, pre-acquisition to post-review gains for Chinese Geography program.

Source	df	MS	F	p
Review	1	2.65		
Step	1	372.30	5.76	<.025*
Ability	2	1043.42	16.15	<.001*
R x S	1	3.63		
R x A	2	111.28	1.72	<.20
S x A	2	15.88		
RxSxA	2	86.48		
ERROR	62	64.60		

---

*Indicates effect accepted as reliable ($p < .05$).

Mean gain for significant main effects on the pre-acquisition to post-review differences analysis of the Geography of China program:

Gain:	Step ($p < .025$)	SS	50.30
		LS	39.16
	Ability ($p < .001$)	HA	59.26
		MA	47.59
		LA	27.34

Table XVIII

Analysis of Variance, post-acquisition to post-review gains for Decimals program.

Source	df	MS	F	p
Review	1	8.04		
Step	1	27.00		
Ability	2	228.82	6.22	<.005*
R x S	1	127.32	3.46	<.10
R x A	2	12.13		
S x A	2	13.48		
RxSxA	2	26.34		
ERROR	59	36.77		

Table XIX

Analysis of Variance, post-acquisition to post-review gains for Squaring program.

Source	df	MS	F	p
Review	1	114.58	3.98	<.10
Step	1	4.92		
Ability	2	516.36	17.94	<.001*
R x S	1	160.97	5.59	<.025*
R x A	2	14.24		
S x A	2	7.99		
RxSxA	2	1.20		
ERROR	62	28.79		

Table XX

Analysis of Variance, post-acquisition to post-review gains for Chinese Geography program.

Source	df	MS	F	p
Review	1	642.99	38.73	<.001*
Step	1	0.79		
Ability	2	0.14		
R x S	1	19.05	1.15	
R x A	2	27.50	1.66	<.20
S x A	2	4.97		
RxSxA	2	25.92	1.56	
ERROR	62	16.60		

---

*Indicates effect accepted as reliable ($p < .05$).

Mean gain for significant effect (p <.005) due to ability level on the post-acquisition to post-review differences analysis of the Decimals program:

Gain: HA 1.83
MA -3.84
LA 11.15

Mean gain for significant effect (p <.001) due to ability level on the post-acquisition to post-review differences analysis of the Squaring program:

Gain: HA -12.87
MA -3.34
LA 9.76

Mean gain for significant effect (p <.001) due to review on the post-acquisition to post-review differences analysis of the Geography of China program:

Gain: IR -6.64
DR 8.00

Table XXI

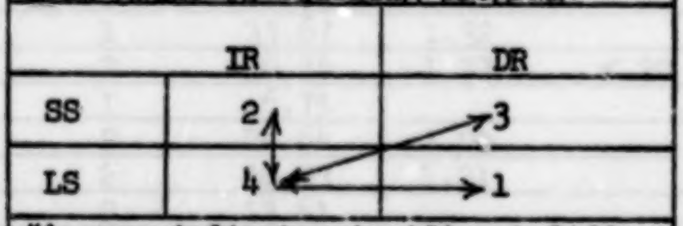
Means and Critical Differences: S x R Interaction, post-review less post-acquisition mean gains on Squaring program. The Critical Difference is 6.20.

Review			IR		DR	
	Step		SS	LS	SS	LS
		Means	-0.94	-9.54	-2.09	3.96
IR	SS	-0.94		-8.60*	-1.05	4.90
	LS	-9.54			-6.45*	13.50*
DR	SS	-2.09				6.05
	LS	3.96				

*Indicates an effect accepted as reliable ($p \leq .05$).

Figure III

Rank Orders: S x A Interaction, post-review less post-acquisition mean gains on Squaring program.



*Arrows indicate significant differences between treatments.

Table XXII

Analysis of Variance, post-acquisition to delayed retention gain for Decimals program.

Source	df	MS	F	p
Review	1	63.20	1.40	
Step	1	80.82	1.80	<.20
Ability	2	16.16		
R x S	1	6.57		
R x A	2	26.25		
S x A	2	46.99	1.04	
RxSxA	2	5.97		
ERROR	59	44.99		

Table XXIII

Analysis of Variance, post-acquisition to delayed retention gains for Squaring program.

Source	df	MS	F	p
Review	1	81.43	2.41	<.20
Step	1	4.76		
Ability	2	128.37	3.80	<.025*
R x S	1	.52		
R x A	2	39.88	1.18	
S x A	2	10.96		
RxSxA	2	19.14		
ERROR	62	33.81		

Table XXIV

Analysis of Variance, post-acquisition to delayed retention gains for Geography of China program.

Source	df	MS	F	p
Review	1	59.91	2.35	<.20
Step	1	33.67	1.32	
Ability	2	50.34	1.97	<.20
R x S	1	6.75		
R x A	2	26.64	1.04	
S x A	2	30.93	1.21	
RxSxA	2	13.61		
ERROR	62	25.53		

---

*Indicates an effect accepted as reliable (p <.05).

Mean gain for significant effect (p <.025) due to ability level on the post-acquisition to delayed-retention differences analysis of the Squaring program:

Gain: HA -8.88
MA -8.45
LA 1.14

Table XXV

Analysis of Variance, post-review to delayed retention gains for Decimals program.

Source	df	MS	F	p
Review	1	94.08	1.24	
Step	1	7.02		
Ability	2	217.82	2.86	<.10
R x S	1	225.78	2.96	<.10
R x A	2	56.34		
S x A	2	113.42	1.49	
RxSxA	2	33.23		
ERROR	59	76.13		

Table XXVI

Analysis of Variance, post-review to delayed retention gains for Squaring program.

Source	df	MS	F	p
Review	1	1.83		
Step	1	0.13		
Ability	2	167.66	5.63	<.01*
R x S	1	170.56	5.73	<.025*
R x A	2	63.48	2.13	<.20
S x A	2	.57		
RxSxA	2	24.60		
ERROR	62	29.76		

Table XXVII

Analysis of Variance, post-review to delayed retention gains for Geography of China program.

Source	df	MS	F	p
Review	1	162.95	5.46	<.025*
Step	1	33.07	1.11	
Ability	2	104.60	3.51	<.05*
R x S	1	6.62		
R x A	2	110.94	3.72	<.05*
S x A	2	4.68		
RxSxA	2	2.70		
ERROR	62	29.83		

---

*Indicates an effect accepted as reliable (p <.05).

Mean gains for significant effect (p <.01) due to ability level on the post-review to delayed-retention differences analysis of the Squaring program:

Gain: HA 4.00
MA -4.82
LA -8.62

- -

Mean gains for significant main effects on the post-review to delayed-retention differences analysis of the Geography of China program:

Gains: Review (p <.025) IR -7.47
DR -14.84

Ability (p <.05) HA -13.53
MA -14.64
LA -5.28

Table XXVIII

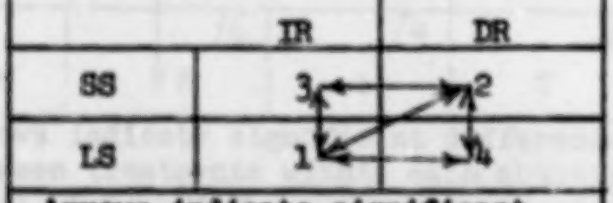
Means and Critical Differences: S x R Interaction, post-review to delayed-retention on the Squaring program. The Critical difference for $p = .05$ in this analysis is 6.30.

Review	Step		IR		DR	
			SS	LS	SS	LS
		Means	-6.63	1.12	0.13	-7.20
IR	SS	-6.63		-7.75*	-6.76*	0.57
	LS	1.12			-0.99	8.32*
DR	SS	0.13				7.33*
	LS	-7.20				

*Indicates an effect accepted as reliable ($p \leq .05$).

Figure IV

Rank Orders: S x R interaction, post-review to delayed-retention on Squaring program.



Arrows indicate significant differences between treatments.

Table XXIX

Means and Critical Differences: R x A Interaction, post-review to delayed-retention for Geography of China program. The critical difference for p = .05 in this analysis is 6.30.

Ability			High		Medium		Low	
	Review		IR	DR	IR	DR	IR	DR
		Means	-4.15	-22.92	-15.64	-13.64	-2.62	-7.96
High	IR	-4.15		18.77*	11.49*	9.49*	-1.53	3.81
	DR	-22.92			-7.28*	-9.28*	-20.30*	-14.96*
Medium	IR	-15.64				-2.00	-13.02*	-7.68*
	DR	-13.64					-11.02*	-5.68
Low	IR	-2.62						5.34
	DR	-7.96						

* indicates an effect accepted as reliable (p ≤ .05)

Figure V

Rank Orders: Delayed-retention less post-review means on Chinese Geography program

	High	Medium	Low
IR	2 ↑ 1	5 2	1 1
DR	6 ↓ 2	4 1	3 2

Arrows indicate significant differences between treatments within each ability level. (←→) The numbers in the large boxes reflect the rank order of the treatments within each ability level. The number enclosed in the upper right corner of each box reflects the position of that cell over the entire interaction. A rank of 1 indicates the best treatment.

Appendix B

Criterion Tests

DELAYED SQUARING TEST

Date ______________________ Name ______________________
Class ______________________

Answer all the questions carefully.

Do the "squaring" problems in the way that the program told you to do them, and then write down the answers in the blank spaces. Don't write the problems down on paper and multiply them out, even if you are also able to do them that way.

*1. The square of 45 is 2025.
2. The square of 2.5 is 6.25.
*3. The square of 65 is 4225.
*4. The square of 15 is 225.
*5. The square of 85 is 7225.
6. The square of .15 is .0225.
*7. The square of 35 is 1225.
8. The square of 8.5 is 72.25.
*9. The square of 75 is 5625.
*10. The square of 95 is 9025.
*11. The square of 55 is 3025.
12. The square of .065 is .004225.
*13. The square of 25 is 625.

*14. When you square a two-place number which ends in 5, you should multiply the " ones -column" number times itself, and then multiply the " tens -column" number times its follower.

15. When you square the number 25, which ends in five, you should multiply five times five, and then multiply two times three.

Name ______________________________

*16. Would the rule which you learned from the booklet help you to square 17 or 31?

a. It would help me square 17, but not 31.

b. It would help me square 31, but not 17.

c. Yes, it would help me square both of them.

[✓ d.] No, it would not help me square either of them.

17. Would the rule which you learned from the booklet help you to square 82 or 52?

a. It would help me square 52 but not 82.

b. Yes, it would help me square both of them.

c. It would help me square 82 but not 52.

[✓ d.] No, it would not help me square either of them.

*18. When you square a two-place number ending in 5, your answer always ends in?

a. The "tens-column" number.

b. The follower.

c. The "tens-column" number times its follower.

[✓ d.] 25.

19. Would the rule which you learned from the booklet help you to square 54, 36, 92, 85, 79, 56, 45?

a. Yes, it would help me square 45, 85, 56, and 54.

[✓ b.] Yes, it would help me square 85 and 45.

c. No, it would not help me square 85.

d. No, it would not help me square any of them.

Name ______________________

20. When you square a two-place number ending in 5, your answer always ends in 26.

a. True ____

b. False ✓

21. When you square a two-place number ending in 5, the tens-column number is multiplied times its follower.

a. Before adding the follower and 5.

b. After multiplying the ones-column number times its follower.

c. Before multiplying the two-column number times one.

[✓ d.] After multiplying the ones-column number times itself.

*22. The follower of a number is always how much larger than the number?

[✓ a.] 1 number larger.

b. 2 numbers larger.

c. Two times as large.

d. It depends on the number.

What is the follower of these numbers?

23. 5 ? 6

24. 2 ? 3

25. 9 ? 10

26. 8 ? 9

Multiply these numbers times their followers:

27. 8 = 72

28. 7 = 56

29. 2 = 6

30. 4 = 20

Name ____________________

*31. What is the word which means "a number multiplied times itself"?

square.

WRITE THE ANSWERS TO THESE PROBLEMS:

*32. 1 times 2 = 2

*33. 4 times 5 = 20

*34. 7 times 8 = 56

*35. 3 times 4 = 12

*36. 6 times 7 = 42

*37. 8 times 9 = 72

*38. 2 times 3 = 6

*39. 9 times 10 = 90

*40. 5 times 6 = 30

41. 25 times 25 = 625

42. .7 times .5 = .35

43. 45 times 4.5 = 202.5

44. 8.5 times 8.5 = 72.25

DELAYED DECIMALS TEST

Date ____________________

Class ____________________ Name ____________________

*1. 9 represents nine (or 9) ones.

2. 004.000 represents four (or 4) ones.

*3. 10 represents ten ones or one (or 1) ten.

4. 3.5 represents three (or 3) ones plus five (or 5) tenths.

*5. 45 represents four tens plus five (or 5) ones.

*6. .621 represents six (or 6) tenths plus two (or 2) hundredths plus one (or 1) thousandths.

*7. 300 represents three (or 3) hundreds.

8. 0601.00 represents six hundreds plus one one.

*9. 509 represents five (or 5) hundreds plus nine (or 9) ones.

10. 00050.0000 represents five (or 5) tens or fifty (or 50).

11. 8296 represents eight thousands plus two (or 2) hundreds plus nine (9) tens plus six (or 6) ones.

12. 151.845 represents one hundred plus five (or 5) tens plus one (1) one plus eight (or 8) tenths plus four (4) hundredths plus five thousandths

WRITE THE FOLLOWING IN DECIMAL FORM. WATCH WHERE YOU PLACE DECIMALS.

13. Four ones 4.

14. Three hundredths .03

15. Two tenths .2

16. Six thousandths .006

17. Four ones plus six tenths 4.6

18. One hundredth .01

19. Eight thousandths .008

20. One ten or ten ones 10.

21. Twelve hundredths .12

22. Two plus five hundredths 2.05

23. Six tens plus seven tenths plus eight hundredths plus four thousandths 60.784

Name ______________________

WRITE THE FOLLOWING COMMON FRACTIONS IN DECIMAL FRACTION FORM (WATCH DECIMAL) (EXAMPLE.... 5/10 = .5)

24. 1/10 = .1
25. 7/10 = .7
26. 1/100 = .01
27. 8/100 = .08
28. 9/10 = .9
29. 5/100 = .05
30. 76/100 = .76
31. 7/1000 = .007
32. 3/1000 = .003
33. 100/100 = 1.
34. 27/1000 = .027
35. 9/1000 = .009
36. 535/1000 = .535
37. 1999/1000 = 1.999

WRITE OUT THE FOLLOWING DECIMAL FRACTIONS IN WORDS. (EXAMPLE.... .08 = eight hundredths)

38. .500 has the value of five tenths
or
fifty hundredths
or
five hundred thousandths
39. .06 has the value of six hundredths
40. .8 has the value of eight tenths
41. .030 has the value of three hundredths
or
thirty thousandths
42. .007 has the value of seven thousandths
*43. .08 has the value of eight (word) hundredths (word)
or
eighty (word) thousandths (word)

Name ____________________

44. .175 has the value of one hundred seventy-five thousandths

WRITE THE FOLLOWING DECIMAL FRACTIONS AS COMMON FRACTIONS. (EXAMPLE... .3 = 3/10)

45. .80 = 8/10 or 80/100
46. .04 = 4/100
47. .1 = 1/10
48. .111 = 111/1000
49. .9 = 9/10
50. .28 = 28/100
51. .063 = 63/1000
52. .070 = 7/100 or 70/1000
*53. Three-tenths is written in numbers as 3/10 or .3 .
54. Fifty-hundredths is written in numbers as 50/100 or .50 .
55. Twenty-one hundredths is written in numbers as 21/100 or .21 .
56. Nine hundred and ninety-eight thousandths is written in numbers as 998/1000 or .998 .
*57. 5/10 is read as five tenths
58. 99/100 is read as ninty - nine hundreadth
59. 8/1000 is read as eight thousandths
*60. 9/1000 is nine divided by one thousand
61. 54/100 is fifty-four divided by one hundred

WRITE AS DECIMAL FRACTIONS:

62. 36 and 9 tenths 36.9
63. 518 and 213 thousandths 518.213
64. 69 and 25 hundredths 69.25
65. 7 and 2 thousandths 7.002
66. 88 and 4 tenths 88.4

Name ____________________

67. 375 and 53 hundredths 375.53

68. 1634 and 1256 ten-thousandths 1634.1256

69. 930 and 784 ten-thousandths 930.0784

70. 2060 and 5 thousandths 2060.005

71. 429 and 1 hundredth 429.01

72. 70 and 6 ten-thousandths 70.0006

73. 539 and 41 thousandths 539.041

74. 2103 and 1 ten-thousandth 2103.0001

75. 6254 and 7 hundredths 6254.07

76. 435 and 9 tenths and five thousandths 435.905

WRITE THE WORDS GREATER THAN OR LESS THAN OR EQUAL TO IN THE BLANK BELOW TO MAKE A TRUE STATEMENT. (EXAMPLE... 2 IS LESS THAN 3) WATCH DECIMAL

77. 1.3 is greater than .13

78. .01 is less than .1

79. .10 is greater than .04

80. .00 is equal to .000

81. 2.8 is greater than 2.0

82. .09 is less than .1

83. .005 is less than .03

84. 1.6 is greater than 1.08

85. .12 is greater than .012

86. .01 is equal to .010

87. 5.78 is greater than 5.69

88. 687 is less than 687.9

DELAYED CHINESE GEOGRAPHY TEST

Date:____________ NAME____________

Class:____________

1. On what continent is China located? Asia

2. What ocean separates the China coast from the western coast of the United States? Pacific

3. What is meant by the word "conterminous"? having the same boundary

4. What is the geographic characteristic of the region of Tibet? b.

PICK ONE:
a. desert
b. Plateau

5. Name the two major political and physical divisions of China.
1. Taiwan (Formosa) 2. Mainland China

6. Why don't the rainbearing winds from the Indian subcontinent reach the regions of Tibet and Sinkiang? the winds are blocked by the Himalaya Mountains

6a. Land which is suitable for farming in China is found mostly in the eastern area.
(western/eastern)

7. Why do the Chinese call Mainland China "Chung Kuo"? because it means Middle Kingdom

8. What region is known as the "roof of the world"? Tibet

9. What is the approximate area of the conterminous United States? three million square miles.

10. The word "Himalaya" means house of snow

10a. Two important mountain ranges in Greater China are (A) the Himalaya Mountains in southern Tibet and (B) the Chin Ling Mountains in China Proper
(what region)?

11. What mountain is the highest in the world? Mt. Everest.

12. What very large country borders China to the north and northwest? Russia

Name__________________

12a. China lies mostly between 25 ° and 50 ° north (north/south) latitude.

13. What does Chung Kuo mean? Middle Kingdom or Central Country.

14. Formosa is the other name for which of these areas? a

pick one:
a. Taiwan
b. Mainland China

15. Greater China is how much larger than the United States? three times larger.

16. Where is the Mongolian Republic located? c

pick one:
a. south-southwest
b. south-southeast
c. north-northwest
d. north-northeast
of China.

17. What two island countries are located off the coast of China? Japan and Philippine Islands.

18. In what range of mountains is the highest mountain in the world located? Himalaya.

19. What political party controls Mainland China? Communist.

20. What subcontinent borders China on the Southwest? Indian

21. The highest mountain in the United States is in Alaska. Its name is Mt. McKinley.

22. Name the three bodies of water which immediately border Mainland China. Yellow Sea, East China Sea, and South China Sea.

23. Hong Kong is a colony of what country? Great Britain

24. Of what political importance is Hong Kong to the other nations of the world? as a listening post.

25. Does Tibet border Mainland China? yes.

26. What is the name of the man made division between China Proper and Inner Mongolia? the Great Wall.

27. There is a group of countries (one country/a group of countries) on the Indian Subcontinent.

Name ______________

28. How many regions make up Greater China? five

28a. Why do only a few people live in the regions of Sinkiang, Tibet, and Inner Mongolia? Rainfall is not sufficient for agriculture; Tibet: cold climate and rugged terrain. Sinkiang: desert

29. What does "Formosa" mean in the language of the Portuguese? beautiful.

30. Name the five neighbors of Chung Kuo (Mainland China) that form the German word MUSIK.

Mongolian Republic

U.S.S.R.

Southeast Asia,

Indian Subcontinent

Korea.

31. What is the geographical description of a basin? broad depression in the Earth's surface.

32. What political party controls Taiwan? Nationalists.

33. Name the region that borders China on the southeast. Southeast Asia.

34. What two countries are larger than China? U.S.S.R. and Canada.

35. Name the two important mountain ranges in China. 1. Chin Ling 2. Himalaya.

36. Which political party of China does the United States support? Nationalist?

37. What countries border China on the northeast? Korea and U.S.S.R..

38. What are the two major deserts in China? Takla Makan and Gobi.

38a. Tell in your own words where the two great deserts in China are located. Takla-Makan-Tarim Basin, Sinkiang.

Gobi - Inner Mongolia

Name ____________________

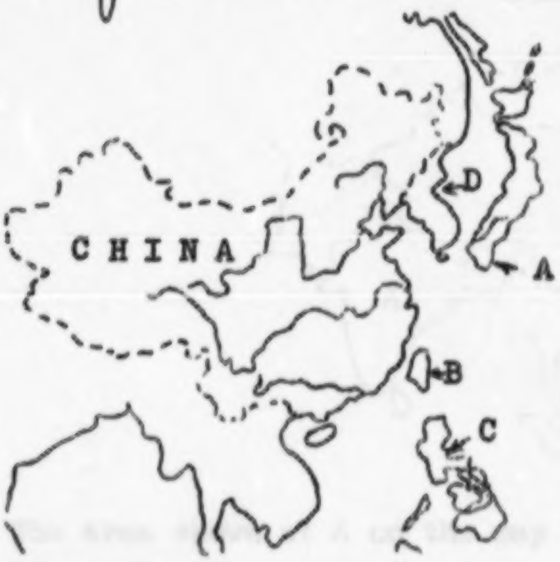
39. Name the areas located at points A, B, C, and D, on the map below.

A. Japan

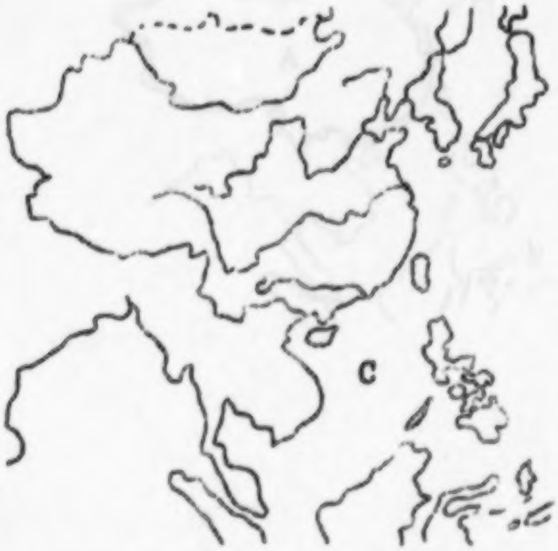
B. Taiwan or Formosa

C. Philippine Republic

D. Korea

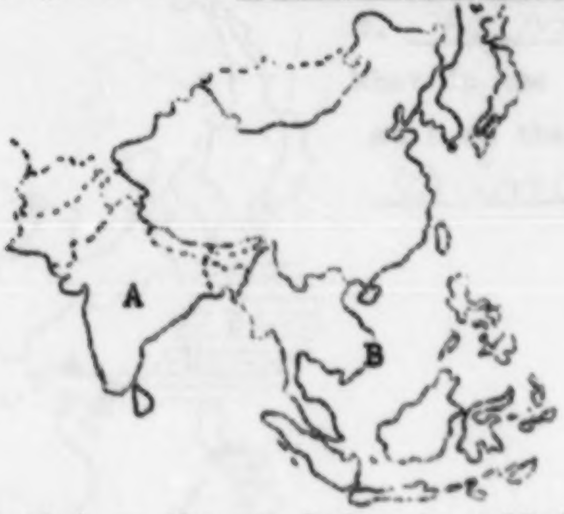


40. At C, on the map below, is the South China Sea. It borders the region southeast of China known as Southeast asia.

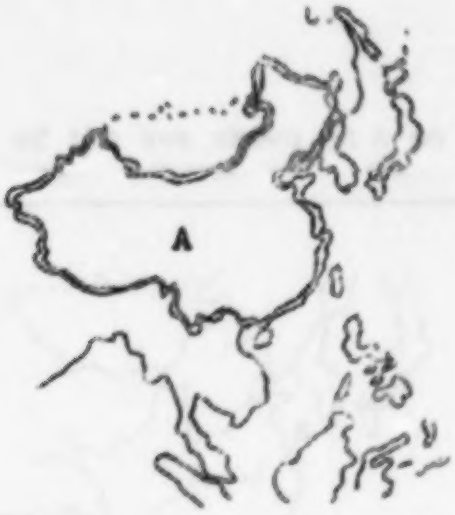


Name: ____________

41. On the map, the area at point A is Indian Subcontinent and the area at point B is Southeast Asia.



42. The area shown at A on the map below is called Chung Kuo by the Chinese.



Name ________________

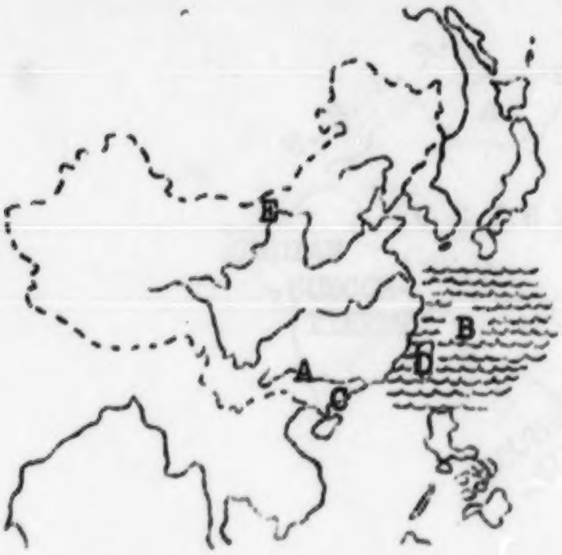
43. What are the names of the two rivers, shown at points A and E on the map below? A. Si Kiang E. Hwang Ho

What are the names of the two areas shown by points C and D?

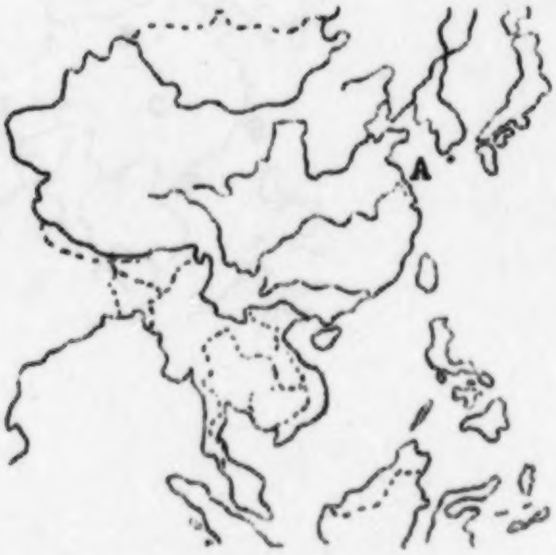
C. Hong Kong

D. Formosa

What is the name of the sea, shown at B on the map? B. East China Sea

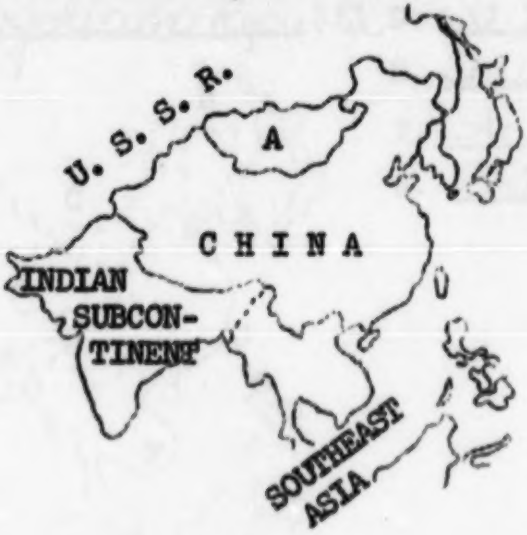


44. What is the name of the sea shown at A on the map? Yellow Sea.



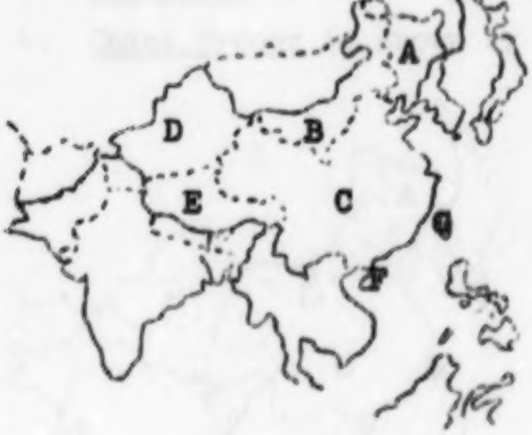
Name ____________________

45. **What is the name of the area shown at A on the map below?** ______ Mongolian Republic.


U. S. S. R.
A
CHINA
INDIAN SUBCONTINENT
SOUTHEAST ASIA

46. **Name the places indicated on the map below.**

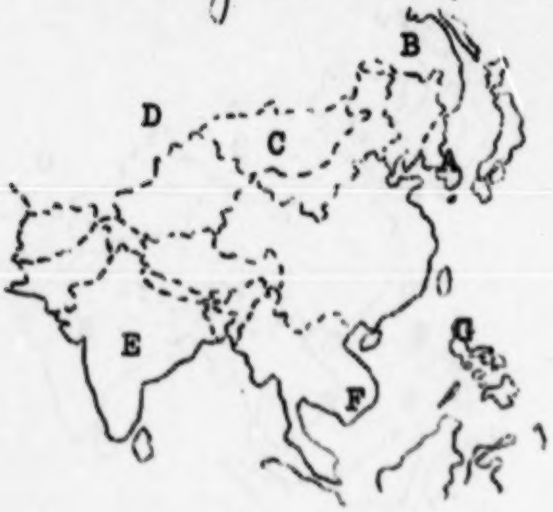
A. Manchuria
B. Inner Mongolia
C. China Proper
D. Sinkiang
E. Tibet
F. Hainan
G. Formosa



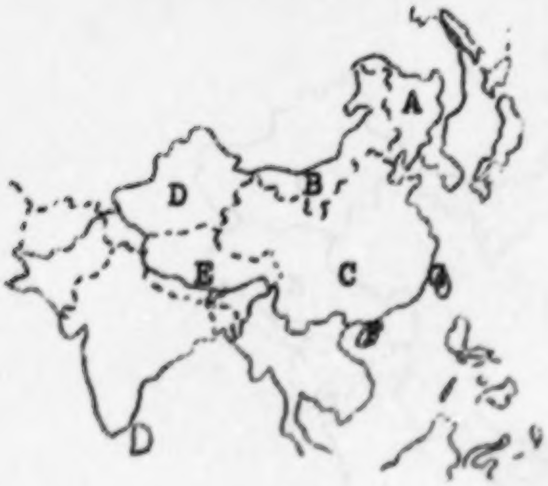
Name ____________________

47. Name the places indicated on the map below.

A. Korea
B. U.S.S.R.
C. Mongolian Republic
D. U.S.S.R.
E. Indian Subcontinent
F. Southeast Asia
G. Philippine Republic



48. 1. **Tibet** is shown by what letter on the map below? E
2. **Inner Mongolia** is shown by what letter on the map below? B
3. **Manchuria** is shown by what letter on the map below? a
4. **China Proper** is shown by what letter on the map below? C
5. **Taiwan or Formosa** is shown by what letter on the map? H
6. **Hainan** is located at what letter on the map? F
7. **Sinkiang** is shown by what letter on the map? D

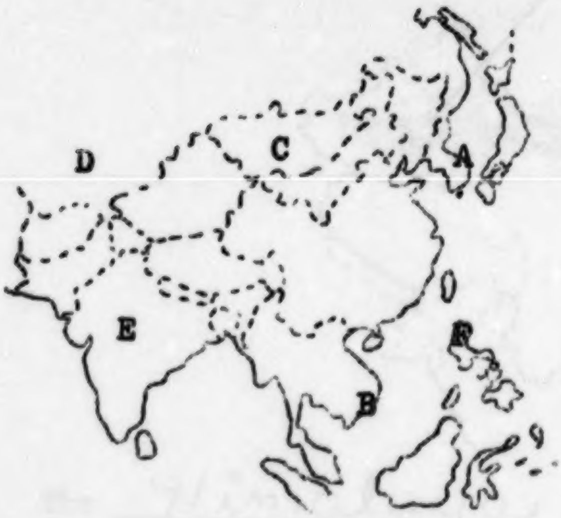


Name ____________________

49. 1. The **U.S.S.R.** is shown by what letter on the map? D

2. The **Indian Subcontinent** is shown by what letter on the map? E

3. **Korea** is shown by what letter on the map? a

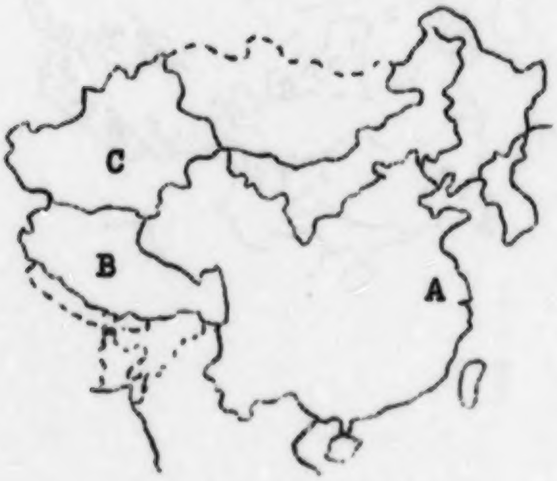


4. The **Mongolian Republic** is shown by what letter on the map? c

5. **Southeast Asia** is shown by what letter on the map? B

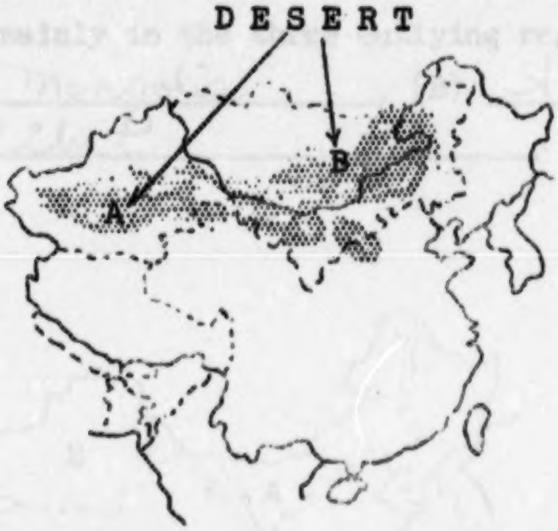
6. The **Philippine Republic** is shown by what letter on the map? F

50. Most of China's farmlands are located at a (letter). (See map below).

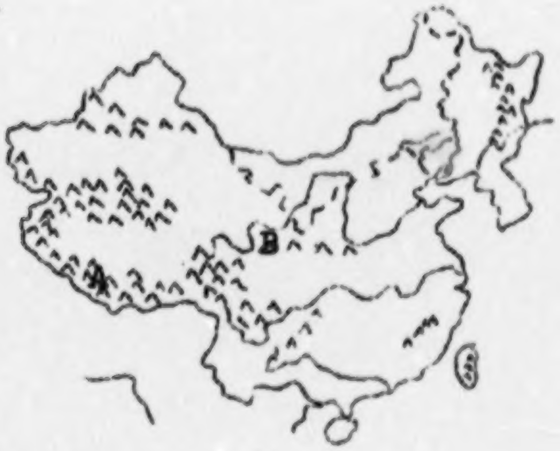


Name ______________

51. 1. The Takla Makan desert is at A.
2. The Gobi desert is at B.



52. The Himalaya Mountains are at A. The Chin Ling Mountains are at B.



Name ________________

53. China is a vast country. However, it contains much unproductive land, found mainly in the three outlying regions of:
(A) Inner Mongolia, (B) Sinkiang
and (C) Tibet.